Imperial College London
Department of Earth Science and Engineering
MSc in Environmental Data Science and Machine Learning

Independent Research Project
Final Project

Modelling Snow Water Equivalent from Snow Depth Measurements Using a LSTM with Multi-Models System

by
Yulin Zhuo

Email: yulin.zhuo22@imperial.ac.uk
GitHub username: edsml-yz6622
Repository: https://github.com/ese-msc-2022/irp-yz6622

Supervisors:
Ms. Niamh French
Dr. Philippa J. Mason
Ms. Corinna Frank

September 2023

Table of Contents

Abstract

In a changing global climate, accurate estimation of Snow Water Equivalent (SWE) becomes more crucial for not only managing water resources, but also mitigating flooding, and understanding climate change impacts. While conventional SWE estimation methods, such as thermodynamic snow models, and empirical regression models (ERMs), show limitations, particularly for transferability, data requirements, and temporal resolutions. As a solution, a novel Multi-model System (MLSTM) combining several Long Short-Term Memory (LSTM) models is proposed, exploiting the advantages of deep learning to capture spatial-temporal snowpack dynamics. The model will use the snow depth and some meteorological variables such as air temperature, and precipitation as input to estimate SWE values in four different geographical countries. In this research, the implementation of multiple models is crucial to enhance the model's transferability. Additionally, adding the snow classification scheme by (Liston & Sturm, 2021) to the list of input features improved the model's transferability significantly. Advanced feature selection by using the *Captum* package by (Kokhlikyan, N., et al. 2020) is applied to mitigate data constraints, thus improving the model's performance. A comparison with the conventional methods, such as iSnobal by (Marks et al., 1999), showed that MLSTM reduces the RMSE from 8 cm to 6.512 cm, demonstrating that MLSTM provides more accurate SWE estimations.

1 Introduction

1.1 Background

As pointed out by (Barnett et al., 2005), the existing climate models consistently forecast near-surface warming due to greenhouse gases, which inevitably affect the hydrological system. Warmer climates mean less winter snowfall and faster snow melting in spring. This leads to a shift in peak river runoff to winter and early spring, rather than in summer and autumn when the water demand is typically highest. One example is the Western United States, where there is a significant reduction in mountain snowpack and a change in steam-flow seasonality. By the mid-21st century, the spring stream flow peak will advance by around one month. This hydrological change shows severe concerns for water availability which impacts those who heavily rely on meltwater for daily needs, and hence emphasises the need for accurate estimation of snowmelt water volumes.

SWE quantifies the total amount of water contained in the snowpack, which is the reason why an accurate estimation of SWE is crucial, especially in places like California where the annual April snowpack water storage is almost twice as large as surface water reservoir storage (Siirila-Woodburn et al., 2021). Accurate estimation of SWE not only aids in water resource management but also in agriculture and flood prevention. Since snowmelt water is used for agriculture and human consumption by approximate one-sixth of the global population (1.2 billion people) (Barnett et al. 2005). Furthermore, if accurate peak SWE values can be provided, it can also aid in early warning for flooding.

1.2 Literature review

Traditionally, SWE is estimated by converting snow depth measurements into snow density estimates. These methods include Thermodynamic Snow Models, ERMs and Semi-empirical Models. Modern thermodynamic snow models focus on mass and energy balance within the ground-snow-atmosphere system but often have high requirements on specific atmospheric variables (Winkler et al., 2021). ERMs, such as the models proposed by (McCreight & Small, 2014), and by (Jonas et al., 2009) strongly rely on the linear relationship between snow depth and SWE. However, the model by (Jonas et al. 2009), may not be able to provide daily resolutions, which makes it unsuitable for certain applications such as water resources management, which requires high-resolution data. The semiempirical models combine theoretical principles and empirical observations. One of the semiempirical models by (Winkler et al., 2021) require snow depth only, but they significantly depend on the initial condition such as density.

Some of these models are more suitable in specific areas, for example, the models by (Sturm et al., 2010) are suited for use in sparsely populated places like the Arctic.

Recently, there has been some promising development in estimating SWE by using deep learning models, such as Artificial Neural Networks (ANNs) (Ntokas et al., 2021). However, the model still has relatively high data requirements, including 'days without snow since the beginning of winter', 'total solid precipitation in the last 10 days', and more. Even though there are open-source websites that provide meteorological data, many of them do not provide these specific measurements. Hence, it will cost lots of labour and resources to continuously monitor and collect these measurements, and many less developed areas might not have the ability to gather these data.

1.3 Objectives

Given the limitations of existing models, this project aims to develop a LSTM model to accurately estimate daily SWE values. LSTM model is a type of recurrent neural network, which is excellent in processing the time series data. My focus is not only to enhance the transferability to different geographies, which means improving the model's ability to be applied across different regions without loss of accuracy, but also reduce the data requirements. Hence, I explored two different LSTM architectures—a single LSTM (SLSTM) model, and a multi-model system (MLSTM) based on different snow classes—to achieve my objectives. Through these methods, I aim to bridge the gaps in conventional estimation methods.

2 Data

2.1 Snow Depth and Meteorological Variables

A robust dataset comprising 35,811 daily records of snow depth (in cm) and SWE (in cm) measurements was employed to develop a LSTM model. This time series data is collected from four diverse countries: Norway, Canada, Switzerland, and the United States of America (USA), as summarised in *Table 1*. These countries have distinct geographical locations, different climates ranging from polar to temperate, and topographies from alpine to plains. The diversity ensures a comprehensive capture of snow characteristics. A summary map of the station data is shown in *Figure 1*—the star markers show the locations of stations.

In Norway, the dataset provided by NVE (Kart | Sildre, n.d.), gathering 11,732 records spans from 2008 to 2021 and is sourced from 5 distinct sites. These sites are geographically distributed across Southern Norway, specifically within the counties of Viken, Vestland, and Agder. The data span Tundra, Maritime and Montane Forest snow classes.

The dataset from Canada, a total of 15,011 records, was sourced from the Aquarius Time-Series database provided by the Government of British Columbia (Data - AQUARIUS WebPortal, n.d.). These measurements cover the period from 2003 to 2022, is collected from 4 sites located in Western coastal and inland mountainous regions of British Columbia. The snow within these regions belongs to the Boreal Forest, Maritime and Montane Forest classes. All measurements in this dataset were acquired using telemetry.

The dataset from Switzerland, with 2,752 records, covering the period from 2015 to 2020, are collected from EnviDat. Measurements were conducted using both GPS (Koch, F., Henkel, P., Appel, F., Mauser, W., Schweizer, J., 2018) and GNSS (Capelli, A., Koch, F., Marty, C., Henkel, P., Schweizer, J, 2020) technologies at two sites located in Eastern Switzerland. These sites exclusively focus on the Tundra snow class. These methods may need to know the vertical distance to the ground in snow-free conditions to accurately determine snow depth.

The USA dataset, which contains 13,467 measurements from 2012 to 2017, is provided by the NASA National Snow and Ice Data Centre Distributed Active Archive Centre (Larson, K. M. and E. E. Small., 2017). These data are from 14 sites located in the Northern and North Central regions of the USA, including snow classes of Boreal Forest and Prairie. Snow depth in the dataset is measured by computing the height difference between the snow surface and the snow-free surface by using GPS, and SWE is calculated by using the values of snow depth and snow density. Because this method relies on the GPS, the measurements could be skewed in some cases, for example, under the presence of sources of interference or extreme weather.

Table 1. Data Sources of the Training Set

Data Source	No. of Daily Sample	No. of Sites	Source
Norway	11732	5 sites	The Norwegian Water Resources and Energy Directorate (NVE), (Kart | Sildre, n.d.)
Canada	15011	4 sites	Aquarius Time-Series database provided by the Government of British Columbia, (Data - AQUARIUS WebPortal, n.d.)
Switzerland	2752	2 sites	EnviDat GPS-derived data, (Koch, F., Henkel, P., Appel, F., Mauser, W., Schweizer, J., 2018) GNSS data, (Capelli, A., Koch, F., Marty, C., Henkel, P., Schweizer, J, 2020)
United States	13467	14 sites	NASA National Snow and Ice Data Centre Distributed Active Archive Centre. (Larson, K. M. and E. E. Small., 2017)

For a comprehensive understanding of the snow conditions and to consider the potential climatic influences on the recorded snow data, meteorological variables were also incorporated into the model. The dataset chosen for this purpose is the ERA5-Land hourly data by (Copernicus Climate Change Service, 2019), spanning from January 1950 to the present.

ERA5-Land is a reanalysis dataset with higher spatial resolution compared to ERA5 (Climate Data Store (CDS), 2023), designed to offer a more detailed view of land variables' evolution over the decades. The principle behind reanalysis is the fusion of observational data from around the globe with model data, which should provide a precise description of past climates.

The daily *Temperature (°C)*, *Precipitation (m), Snowfall (m of water equivalent),* and *Solar radiation* ($J\,m^{-2}$) are retrieved to be meteorological features. The *Temperature* is crucial as it directly influences snowmelt and accumulation processes. *Precipitation* gives insights into potential snow accumulation. *Snowfall* provides a more direct measure of snow, while *net surface solar radiation* can affect the snow-melting rates. A detailed description of meteorological variables is available in *Appendix 1. Table 1*.

2.2 Snow Classification

The study utilises the Global Seasonal-Snow Classification data set including tundra, boreal forest, maritime, ephemeral (including areas with no snow), prairie, montane forest, ice, and ocean (Liston & Sturm, 2021). The snow classification map used in this study is shown in *Figure 1*. The classes are based on physical attributes such as air temperature, precipitation, and wind speed. The data set consists of a 39-year period from 1981 to 2019, with each snow classification representing an average over this period. The study chose a classification scheme with global coverage at a resolution of around 50 km.

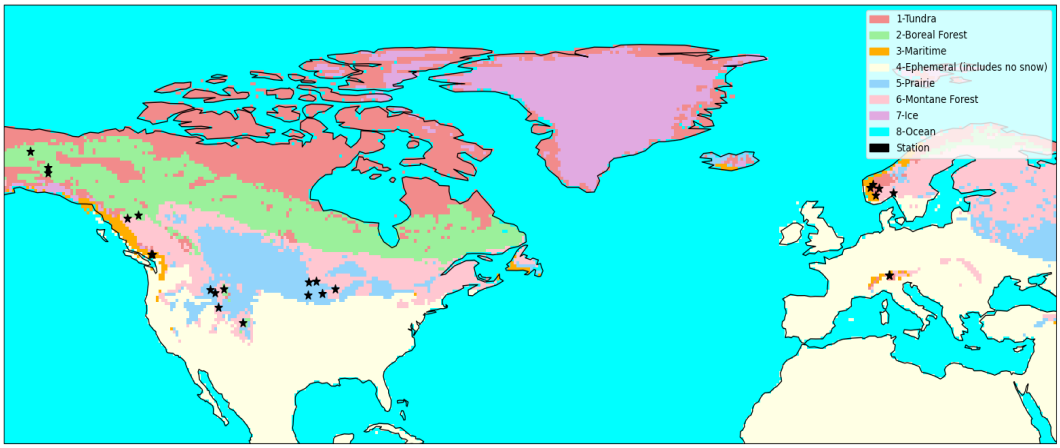

Figure 1. A map showing Global Snow Classification and the locations of stations from USA to Europe.

Since the classification is based on a 39-year average, hence the differences in air temperature and precipitation can cause snow classes in any given year to be different dramatically. Furthermore, the 39-year land cover data was not available for deriving classification, the land cover data used for the classification is from 2018, which may not accurately represent the land cover conditions throughout the entire 39-year period. In addition, the high-resolution map may not perfectly capture the high spatial variability in mountain snow covers, which can be impacted by several factors such as wind directions and topography.

In our dataset, class 4 (Ephemeral, including no snow), class 7 (Ice), and class 8 (Ocean) are not included in this dataset due to dataset limitations. As a result, these four classes will not be included in the multi-mode system. Consequently, only five snow classes: Tundra, Boreal Forest, Maritime, Prairie, Montane Forest are used in the model training. Overall, while the classification data set provides a useful global picture of snow classes, the actual conditions in specific areas and specific years may differ.

2.3 Data Analysis

The training data has a notable presence of low-valued observations as shown in *Figure 1* of *Appendix 2,* primarily spanning the spring to summer periods. This observation aligns with the expected seasonal characteristics. However, to achieve a more balanced dataset and reduce potential biases during model training, some of these summer records with values equal to 0 have been selectively removed.

Table 2. Descriptive statistics for training data

Variables	**Snow depth (cm)**	**Snow water equivalent (cm)**	**Precipitation (m)**	**Snowfall (m of water equivalent)**	**Solar radiation (J m-2)**	**Temperature (°C)**
Count	35810	**35691**	35810	35810	35810	35810
Mean	60.30	20.65	1.83×10^{-3}	1.14×10^{-3}	4.21×10^{6}	2.47
min	0.00	0.00	0.00	0.00	0.00	**-34.77**
25%	0.35	0.020	9.31×10^{-7}	0.00	4.10×10^{5}	-4.63
50%	23.00	6.00	1.53×10^{-4}	3.73×10^{-9}	1.96×10^{6}	1.13
75%	86.87	24.49	1.58×10^{-3}	4.397×10^{-4}	5.93×10^{6}	8.69
Max	555.71	242.91	5.09×10^{-2}	5.08×10^{-2}	$\mathbf{2.23 \times 10^{7}}$	41.19

To get a comprehensive understanding of the data and its characteristics, descriptive statistics (*Table 2*) were generated for each feature in the dataset. These statistics provide insights into the distribution, variability, and central tendency of each feature. It was observed that there were some missing values (in SWE) as shown in *Table 2.* Hence linear interpolation is applied to avoid this. Importantly, the time

interval is also considered when doing the linear interpolation so that it can preserve the time characteristic in time series data.

Given the considerable differences in magnitude across the input features — especially with variables like *Solar radiation* having values in the millions, it becomes crucial to normalise the data to ensure model stability during training.

Except for *Temperature*, all the variables used Min-Max scaling to transform their values to lie between 0 and 1. Specifically, each value was scaled by:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}},$$

$where\ x_{min}\ and\ x_{max}\ are\ the\ minimum\ and\ maximum\ values\ of\ variables\ respectively.$

While *Temperature* uses standardisation scaling where each value was scaled by:

$$x_{scaled} = \frac{x - \mu}{\sigma},$$

$where\ \mu\ and\ \sigma\ represent\ the\ mean\ and\ the\ standard\ deviation\ respectively.$

The choice of scaling methods was driven by the nature of the data. Most features had values greater than zero, while the temperature feature contained both positive and negative values as its nature. Using Min-Max scaling for temperature would result in the compression of its range, causing a loss in the intrinsic variability and patterns within the temperature data. Specifically, the natural zero point and the symmetry around this zero for temperature data would be disrupted, potentially affecting the model's ability to capture temperature-driven dynamics in snow measurements. To preserve the original pattern of the temperature data, standardisation scaling was chosen. The scaling approaches ensure that each feature has a consistent influence on the learning process, preventing any single feature from disproportionately dominating due to its scale.

3 Methods

This project proposes a Long Short-Term Memory (LSTM) model to accurately estimate daily SWE across a range of geographic areas. The choice of LSTM is mainly because of the strengths of having unique abilities to process time-series data by remembering patterns over time, which can capture the temporal evolution of snowpack dynamics. This project followed a well-structured approach including data pre-processing, modelling, and model evaluation. The subsequent sections give details of each stage. The overall stages are shown in *Figure 2*.

3.1 Data Pre-processing

The initial focus was on data from Norway, Canada, Switzerland, and the USA, primarily using the station data from NVE (Kart | Sildre, n.d.), Aquarius Time-Series database (Data - AQUARIUS WebPortal, n.d.) provided by the Government of British Columbia, EnviDat (Koch, F., Henkel, P., Appel, F., Mauser, W., Schweizer, J., 2018)(Capelli, A., Koch, F., Marty, C., Henkel, P., Schweizer, J, 2020), and NASA National Snow and Ice Data Centre Distributed Active Archive Centre (Larson, K. M. and E. E. Small., n.d.). This data will be pre-processed for training. After interpolating the missing values (as detailed in Section 2.3), perturbations will be applied to reflect uncertainty as outlined in the WMO's guidelines: if the measured snow depth is less than 20 cm, the error should not be more than ±1 cm, while if it the snow depth is more than or equal to 20 cm, the error should be around ±5% (Organization, 2017). After applying the perturbations, the data will be scaled as described in Section 2.3. Moreover, the snow classification scheme introduced by (Liston & Sturm, 2021) will be employed in MLSTM training to improve model performance. Regarding to the data set division, the training, validation, and testing set are split in 6:2:2 to ensure that the model have enough data for training while having validation and testing data for hyperparameter tuning and evaluation.

3.2 Modelling

3.2.1 Model Construction

The modelling part consists of creating one LSTM model and one LSTM models system:

1. **Single LSTM model (SLSTM)**: The first model leverages the entire dataset without any classification.
2. **Multi-LSTM models system (MLSTM):** In this system, snow is categorised into five classes according to the snow classification scheme (Liston & Sturm, 2021). Specifically, a distinct LSTM model is designed for each snow class, with the snow class itself being incorporated as one of the inputs. The model performance will be reported as average performance over the snow classes.

3.2.2 Hyper-parameter Tuning and Architecture Adjustment

All models undergo a process of hyper-parameters tuning and architecture adjustment to optimise performance, which involves:

1. **Model architectures**: Using *Grid Search* to evaluate various LSTM architectures with 1-4 layers. *Grid Search* is a method for tuning hyper-parameters, where the model is trained for all possible combinations of parameter values to find the one that can have the best performance.
2. **Number of neurons**: Use *Grid Search* to tune the number of neurons in each layer.
3. **Number of Epochs**: Utilise *Early Stopping* technique to determine the optimal number of epochs, the main idea is to break the training process when there is no improvement observed in the validation loss over a certain number of epochs.

3.2.3 Meteorological Variables and Time Sequence Length Selection

A model incorporating meteorological data, such as *Precipitation*, *Temperature*, and *Solar radiation* from European Centre for Medium-Range Weather Forecasts (Copernicus Climate Change Service, 2019) will be developed. The selections of meteorological features and the time sequence length rely on:

1. Comparison between the *Pearson* and *Spearman Correlation*. *Pearson Correlation* can analyse the linear relationship between two continuous variables, while *Spearman Correlation* can evaluate the monotonic relationships, which can show the non-linear relationship that might be missed by Pearson Correlation.
2. Utilising '*Captum'* tool to compare the features' importance within the hidden layers by looking at their weights. *Captum* (Kokhlikyan, N., et al. 2020) is an interpretability tool for deep learning models, providing researchers an insight to better understand and visualise the significance of features on predictions. Features are ranked based on the weights of them in the hidden layers.
3. Sensitivity Analysis: An iterative process is applied where features are successively removed according to their importance ranks. It helps determine whether the removal of specific elements results in significant changes in performance.

After using *'Captum'* for feature selection, the model would undergo a time sequence length selection, where '*Captum'* tool is also applied to compare the importance of different lengths of time sequences. After feature selection and time sequence length selection, our finding suggested that adding variables may improve the performance, but sometimes not significantly. Given this randomness in model training processes, a T-test is applied to determine if the enhancements yield statistically significant improvements. T-test is a statistical method to see if the difference between two sample groups (in this case, model performances) is because of the randomness by comparing the means of these groups. By following the conventional applications, the threshold of the p-value is set to be 0.05 for testing statistical significance, as recommended by (Dahiru, 2008). The p-value lower than 0.05 indicates a statistically significant improvement in the modified model.

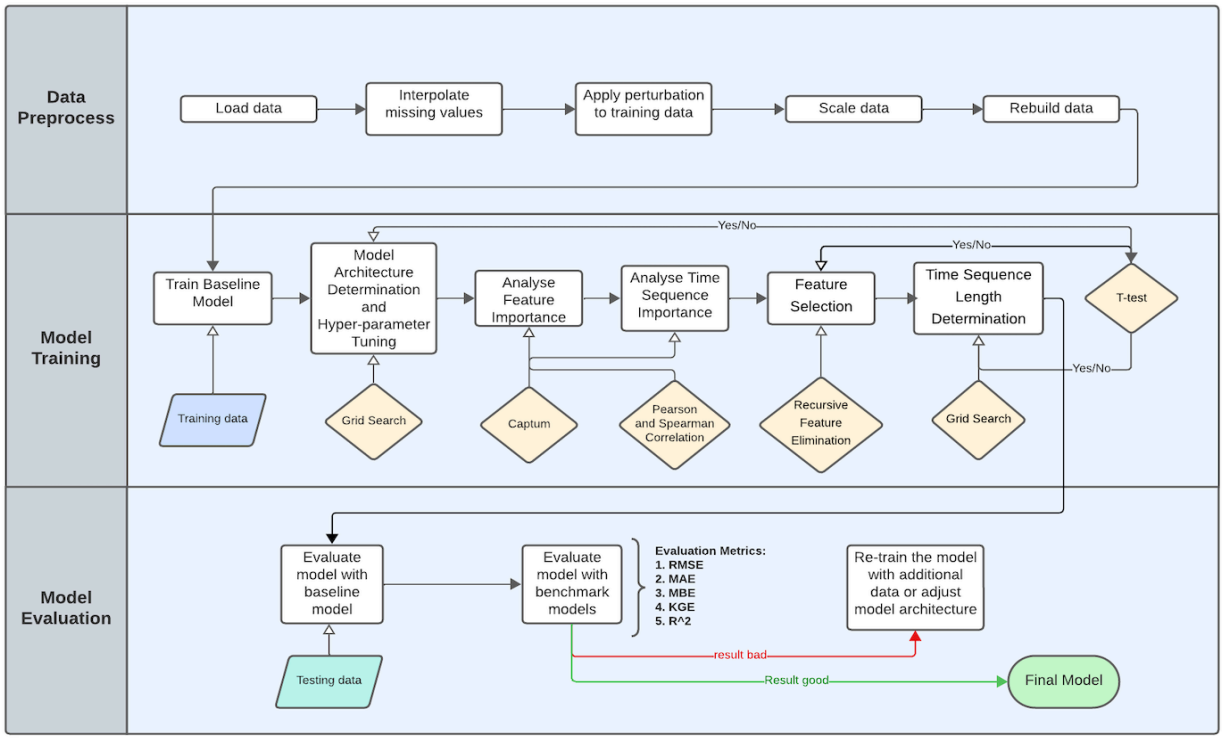

Figure 2. The workflow of model training

3.3 Model Evaluation Metrics

Regarding the performance evaluation, both SLSTM and MLSTM are evaluated by the following performance metrics:

1. Root Mean Squared Error (RMSE): RMSE calculates the square root of the difference between predicted and true values. The lower RMSE score suggests less difference in the predicted and true values. The formula is shown below:

   $RMSE = \frac{\sqrt{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}}{\mathrm{N}}$, where y_i is the true values and $\hat{y}_i$ is the predicted values.

2. Mean Absolute Error (MAE): MAE calculates the absolute value of the difference between predicted and true values. The lower MAE score indicates the predictions are closer to the true values. The formula is:

   $MAE = \frac{\sum_{i=1}^{N}|\hat{y}_i - y_i|}{\mathrm{N}}$, where y_i is the true values and $\hat{y}_i$ is the predicted values.

3. Mean Bias Error (MBE): MBE calculates the average difference between the predicted and true values. The closer the MBE score to zero, the higher accuracy of predictions. A positive MBE indicates an overestimation, while a negative MBE means an underestimation. The formula is presented as:

   $MBE = \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)}{\mathrm{N}}$, where y_i is the true values and $\hat{y}_i$ is the predicted values.

4. Kling-Gupta Efficiency: KGE is a comprehensive hydrological metric, which evaluates the similarity between simulated values and true time-serires by considering their correlation, the ratio of mean, and the ratio of standard deviation. A KGE score of 1 indicates the perfect simulation. The formula is:

$KGE = \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2}$, where r is the Pearson correlation between the predicted and observed values, β is the ratio of standard deviations of predicted and observed values, γ is the ratios of their means.

5. Coefficient of Determination (R^2): R^2 indicates the models' fitness, illustrating the linear relationship between observed values and predicted values. A score close to 1 indicates that the predictions closely match the observations. The formula is defined as:

   $R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}$, where y_i is the true values, $\hat{y}_i$ is the predicted values, and $\bar{y}_i$ is the mean of true values.

RMSE, MAE, and MBE will serve as the key metrics to evaluate the model, combined with KGE and R^2 to give a more comprehensive evaluation, especially for hydrological and regression analysis.

4 Results

4.1 Optimisation of the model setup

In this section, we discuss how different model architectures, hyper-parameters and different feature selections impact the training.

4.1.1 Architectures of the LSTM

Even though testing and experimentation are crucial for selecting a model architecture, we cannot always pursue the model that performs the best in these experiments. A model that is 'good enough' can be more practical — the one that is quicker to train, more interpretable, and demands fewer resources. From the grid search results applied to different LSTM architectures, the findings suggest a worse performance improvement with the addition of further layers. The experiments evaluated multiple models' architecture; the results are shown in *Figure 3*. To clearly explaining the selection process, the model for Maritime is chosen to present for brevity and clarity. Detailed explanations for other models are available in *Appendix 4*. From the green line in *Figure 3*, it can be observed that the model with 3 layers and hidden neurons of 70, 50 and 30 respectively has relatively better performances on RMSE, MAE, and R^2, suggesting a better overall predictive accuracy. Despite slightly higher MBE and lower KGE values indicating the inconsistency in the distribution of the true values and the predicted values, the overall predictions are more accurate than the others on average, as evidenced by the other metrics.

Regarding the training process, the *Early Stopping* technique is implemented to ensure efficiency in epoch determination and overfitting prevention. The *Adam* optimiser was chosen for its efficiency, combined with the decay learning rate to ensure convergence. Given the nature of the target value, the *Rectified Linear Unit (ReLU)* activation function was selected to ensure that the predicted SWE values remain non-negative. This activation function can get the input directly if it is positive, while it outputs zero if it is negative, which makes it suitable for predicting SWE values since it should not be negative. In addition, the large SWE values in the dataset, especially those near peak levels, are sparse. However, it is crucial to accurately estimate those large SWE values in applications. Hence *Mean Squared Error (MSE)* is employed as a loss function to enhance the impact of large errors. *MSE* squares the error so that it makes the loss more sensitive to the larger errors, especially for those large, but sparse SWE values. This ensures that the model can focus on accurately estimating those important points.
Overall, the detailed setup for the models is presented in *Table 1 of Appendix 3*.

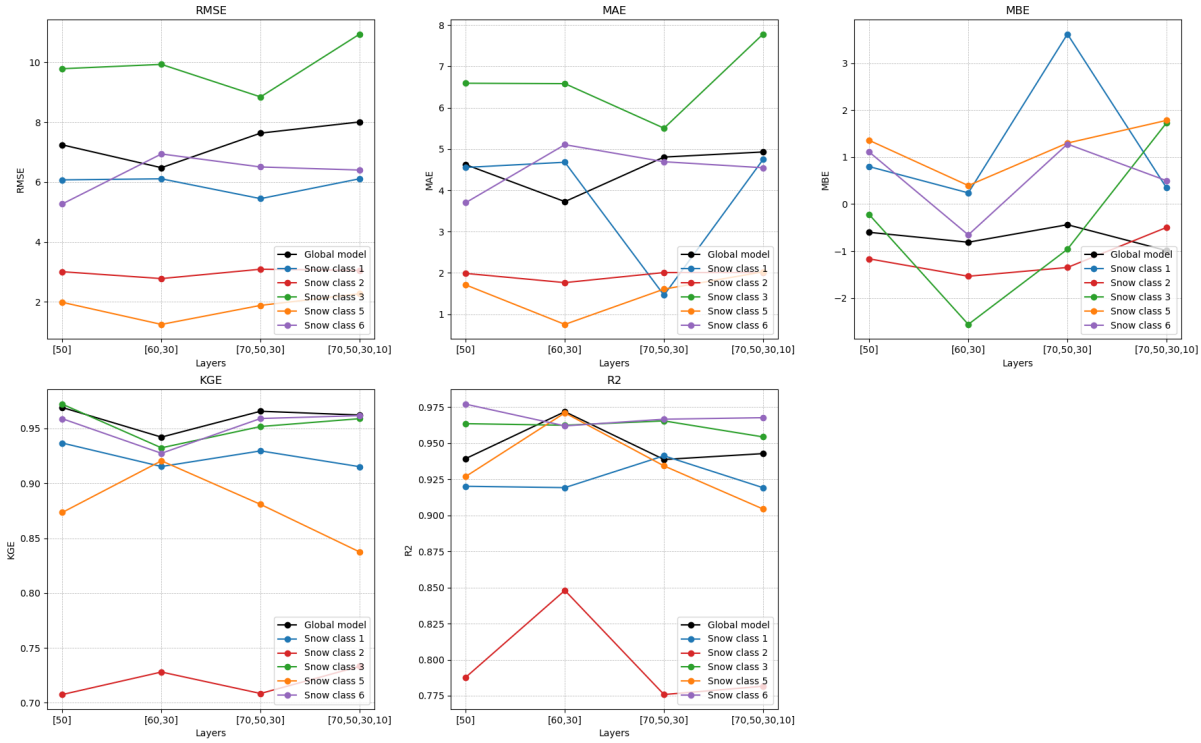

Figure 3. Performance Metrics Results for Grid Search on Different LSTM Architectures

4.1.2 Impact of Features and Sequence Length

To explain the process of feature and time sequence length selection, the results of a representative model (specifically for Maritime) are chosen to present for brevity and clarity. The results and explanations for other models are available in *Appendix 5*.

Feature selection plays an essential role in improving model performance and training efficiency by eliminating irrelevant information. My goal in feature selection is not only to optimise the model's performance but also to reduce the data requirements. For a comprehensive understanding of the significance of different features within the hidden layers, I conducted an overall analysis by employing the '*Captum'* technique. Accordingly, starting with the least important features, I iteratively removed the features in order of their importance and trained the models to compare the performances. As observed in *Figure 4(a), solar radiation* was the least significant feature, hence we removed it from the subsequent model training. Then we repeated this step iteratively. The results of the sensitivity analysis are presented in *Table 3(a).* The model with the combination of *Snow depth, Temperature, and Precipitation* demonstrates the optimal result. *Snow depth* and *Temperature* have a relatively high correlation with *SWE as seen in Table 3(b)*, while *Precipitation* does not. This suggests that *Precipitation* might have a complex non-linear relationship with *SWE* or could interact with *Temperature* to enhance the estimation. Furthermore, in case the two model performances did not show

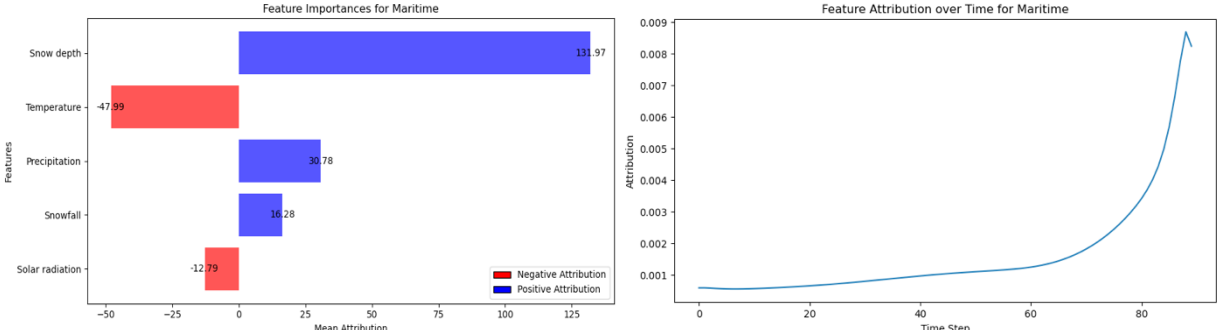

Figure 4. (a) Feature Attributes for Maritime over all samples and all the time steps, (b) Snow depth Attribute over 90 days

Table 3. (a) Performance evaluation of MLSTM models for Maritime with different meteorological features, (b) Pearson and Spearman correlation between all meteorological features and SWE for Maritime

Meteorological Features	**RMSE (cm)**	**MAE (cm)**	**MBE (cm)**	**KEG**	**R^2**
Snow depth, Temperature, Precipitation, Snowfall	14.262	8.687	-1.610	0.934	0.910
Snow depth, Temperature, Precipitation	8.843	5.504	-0.962	0.951	0.965
Snow depth, Temperature	10.335	6.205	-1.344	0.950	0.953

Variables	**Pearson correlation to SWE**	**Spearman correlation to SWE**
Snow depth	0.893031	0.837892
Precipitation	0.040040	0.026125
Snowfall	0.096426	0.205242
Solar radiation	0.011453	-0.172987
Temperature	-0.234563	-0.369044

a large difference, the T-test is applied to statistically evaluate the performances. The results of final feature selections for each model are presented in *Table 3 of Appendix 3.*

In addition to the feature selection, this methodology was also applied in determining the significance of the time sequences. Choosing the right time sequence length is critical to capture key temporal trends. Removing the redundant temporal information helps to improve not only the model's predictive ability but also computational efficiency. By visualising the weights of each day for the most important features (*Snow depth* in most cases), I was able to figure out the time sequence that can have the most significant influence on the target value. In this case, the attribute of *Snow depth* gets increasingly lower as time passes. Since 90% of *Snow depth*'s attributions arise from day 40, as shown in *Figure 4(b).*

The grid search is utilised to methodologically evaluate different time sequences, including 10-day, 20-day, 30-day, 40-day, and 50-day, the performance results are shown in *Figure 5*. Finally, 30-day was selected since it shows strong performance on RMSE, MAE, and R^2 at 30-day. While MBE and KGE did not achieve the optimal result, they have second-place rankings, showing an overall excellent performance. In conclusion, the methodology for feature and time sequence selection has provided detailed insights into optimising the performance of the LSTM models. By focusing on the most important meteorological features and the relevant time sequences, I can ensure efficient and accurate predictions.

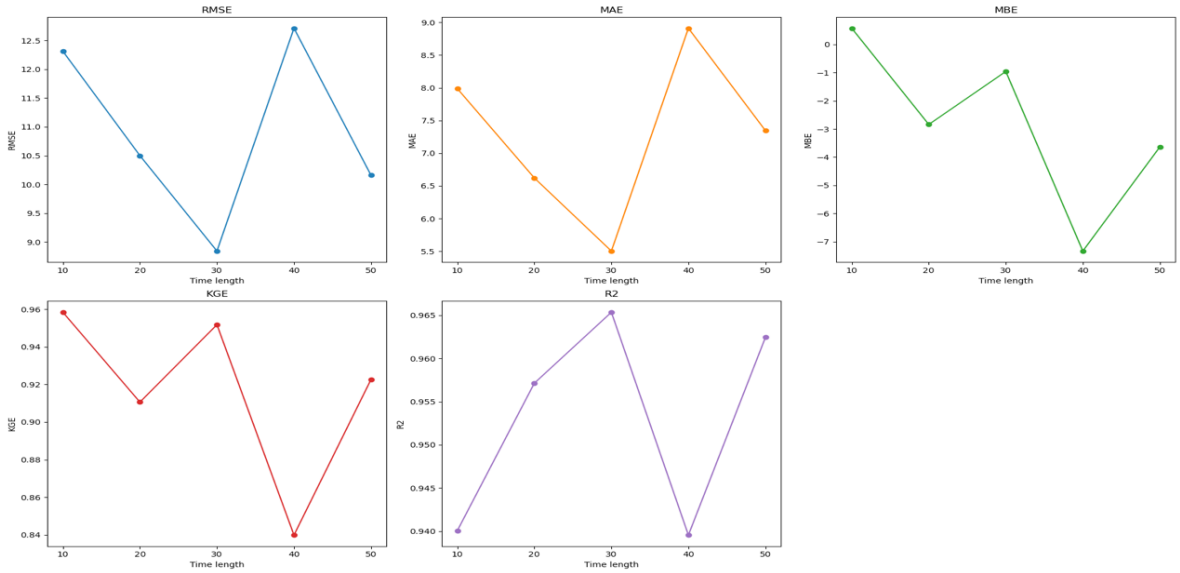

Figure 5. Performance Metrics Results of Grid Search on different time sequence length for Maritime model

4.2 Models performance on the testing set

Regarding the model performance evaluation, MAE, RMSE, and MBE will be used as the key error metrics to evaluate the model performance, as used by (Ntokas et al., 2021). Model performance will be compared with the benchmark model - the MLP model by (Ntokas et al., 2021) and iSnobal (Marks et al., 1999). In addition, Kling-Gupta Efficiency (KGE), and coefficient of determination (R^2) were employed to have a more comprehensive performance analysis of the models.
However, the datasets utilised in this project differed from those in the benchmark models. This difference implies the potential variations in outcomes, which need a more careful and more critical interpretation and comparisons. Additionally, (Ntokas et al., 2021) based their research on the dataset from Canada. And a portion of the data used in this project are also from Canada, highlighting a similarity among the differences.

Table 4. Comparison of performance evaluation metrics of SLSTM and MLSTM with two MLP models (Ntokas et al., 2021) and iSnobal model in Mill-d station (Marks et al., 1999).

	SLSTM	**MLSTM**	**Single MLP ensemble model (SMLP)**	**Multiple MLP ensemble model (MMLP)**	**iSnobal in Mill-D station**
MAE (cm)	3.816	3.679	32.8	29.3	/
RMSE (cm)	6.837	6.512	61.0	51.5	8.0
MBE (cm)	-2.060	-0.049	0.4	0.6	-5.0
KGE	0.895	0.969	/	/	/
R^2	0.969	0.970	/	/	/

From the results presented in *Table 4*, the following observation can be drawn:

1. MAE and RMSE of SLSTM and MLSTM models show a significant improvement compared to the MLP models introduced by (Ntokas et al., 2021) and iSnobal model by (Marks et al., 1999). Specifically, the RMSE are reduced from 61.0 and 51.5 to 6.837 and 6.512, which is an approximately 86.7% to 89.3% decrease.
2. The MLSTM model has a low bias with an MBE value of -0.049, which is 58.5% and 87.75% lower than SMLP and MMLP. However, SLSTM is more biased compared to the MLP models (Ntokas et al., 2021), but it is less biased than iSnobal (Marks et al., 1999).
3. RMSE and MAE values of MLSTM is 4.68% and 3.59% lower than that of SLSTM. MBE of MLSTM is 76.21% lower compared to SLSTM, reduced from -2.060 cm to -0.049 cm, showing that MLSTM is outperforming SLSTM on all these metrics.
4. MLSTM has a better KGE score than SLSTM, which indicates the estimated values from MLSTM are closer to the observed values on the testing set.
5. The R^2 score of SLSTM and MLSTM are close, meaning a robust fit to the data for both models.

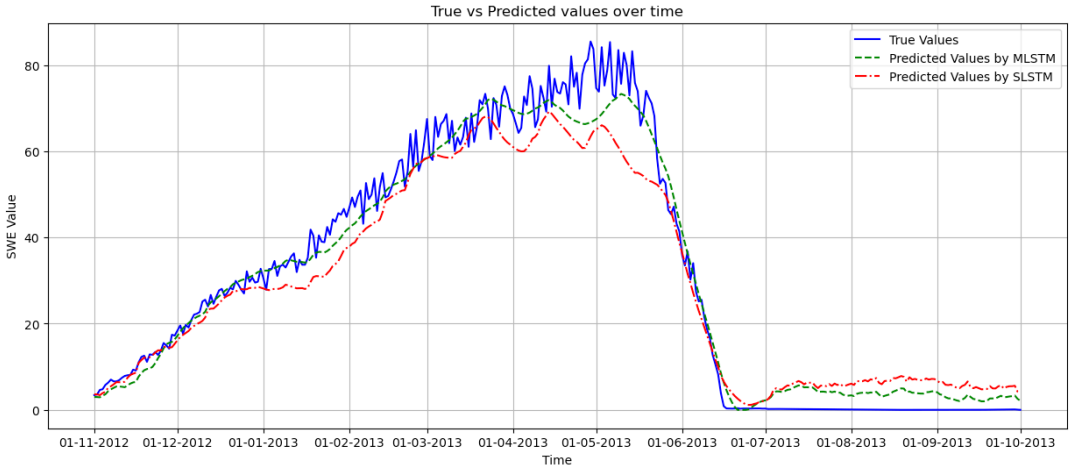

Figure 6. Comparison of True and Predicted Values of MLSTM and SLSTM for the 'canada_4B16P' Station Over a Year (Nov 2012 - Oct 2013)

Figure 6 displays a line plot comparing the observed values and predicted values from SLSTM and MLSTM models at the 'canada_4B16P' station, which located in the Montane Forest regions of Canada, where MLP models by (Ntokas et al., 2021) are evaluated. As observed from the figure, the MLSTM predictions (represented by green line) are generally more aligned with the observed values than those from SLSTM (red line). Additionally, the model's predictions demonstrate a consistent and stable trend while the observed data exhibits a large fluctuation. This not only indicates the model's robustness to capture the trend of data but also shows no obvious overfitting patterns. This result emphasises the MLSTM model's reliability and accuracy.

5 Discussion

Overall, the conclusion is that the MLSTM has a stronger predictive ability than the SLSTM. A potential reason for this might be that the SLSTM is trained on the entire data set, which means it captures the variations in snow conditions at a worldwide scale. This could compromise its performance in specific regions since the snow conditions in different regions can be diverse. The differences in results can be significant. On the other hand, MLSTM is trained based on different snow classes, enabling it to capture the unique characteristics of each snow class better than MLSTM.

Furthermore, within our MLSTM models, only the Maritime model requires precipitation and temperature as input variables. This is potentially because of the nature of Maritime snow, which is generally a warmer type of snow cover typically found in humid coastal regions (Sturm et al., 1995). Even slight temperature fluctuations can have a significant impact on snowpack. Moreover, maritime regions typically receive more precipitation compared to other areas, amplifying the effect of this factor on the snow. As a result, compared to other snow types, Maritime snow is more sensitive to changes in these two features, leading to variations in SWE.

5.1 Challenges

One of the challenges in this project was data pre-processing. Precise measurements of SWE in the real world have considerable challenges. One of the reasons is the inconsistency in data collection methods across different stations, because of my aim of estimating SWE across global regions. Different stations might use various equipment to collect data, this might raise the bias in the readings. In addition, many stations have many missing values, which took me lots of time doing the data filtering and interpolating. If missing values were not addressed well, the model would get confused by the information gaps or even generate null estimations.

5.2 Limitations

MLSTM has its constraints in limiting the application within the specific snow classes: Tundra, Boreal Forest, Maritime, Prairie, and Montane Forest. It lacks the ability to estimate for Ephemeral, Ice, and Ocean due to the absence of data. Besides, one significant constraint of MLSTM is its input requirements for users to know and provide the corresponding snow class of input data when estimating SWE. This requires an initial understanding, which may not always be feasible. Furthermore, different models within MLSTM demand different features, resulting in different data preparation needs. For example, the model for Maritime requires temperature and precipitation data, which is a requirement not shared with the other snow class models (Snow depth is the only requirement for the other models). Another notable limitation is that MLSTM highly relies on snow depth data, requiring continuous data spanning up to 30 days. This requirement can be highly restrictive in regions where continuous data collection might be challenging due to environmental or equipment reasons.

5.3 Future work

Future work can explore more station data covering the entire snow classes. This would potentially improve the model's accuracy across diverse regions. In addition, experiments with different MLSTM

model architectures can be conducted. For instance, training all the models separately on the snow depth only can make the data requirement more consistent. In addition, instead of depending on a single model's output, developing an ensemble model which take average estimations from multiple models can be conducted. This can potentially reduce the bias of individual models to result in more stable and accurate estimations. The meteorological variables used in this project are *Temperature*, *Precipitation*, *Solar radiation,* and *Snowfall*. People can also explore more meteorological variables, such as wind speed, to seek more accurate estimations. In addition, some transformer-based models can be explored, such as BERT. While BERT is a natural language processing model, it can also be adapted for time series forecasting.

6 Conclusion

This project developed and evaluated multiple LSTM models to predict SWE on a global scale. Grid search, '*Captum'* tool, and sensitivity analysis are employed to help in determine the architecture of LSTM. To evaluate the models more comprehensively, I used the testing data which covers different stations in Canada, Norway, Switzerland, and the USA. Its adaptability to diverse datasets emphasises its potential to enhance SWE estimations. Therefore, it can be verified that for the currently available data, the MLSTM model based on snow classes performs better than SLSTM with lower RMSE, MBE, MAE and higher KGE and R^2 score. This shows it potential for accurately predicting SWE globally. In addition, the sensitivity analysis showed that snow depth is the most significant input feature in most cases. The only model that would have a statistically significant improvement after adding *Temperature* and *Precipitation* as input is the MLSTM models for Maritime. In practical applications, MLSTM stands out for its accuracy and precision, especially in region-specific situations. For the regions where the availability of temperature, precipitation, or snow class is limited, SLSTM should be a more suitable choice, which only requires snow depth as input.

Acknowledgments

I believe I am writing these words with a heart full of gratitude. Growing up, I was never the smartest, the most talented, or even the one who worked the hardest. Yet reflecting upon it, I think the reason why I can be here today writing down these words is probably because of the countless hours I have spent striving in unnoticed corners. When I received the offer from IC, I felt as if I had trudged through life unexcitedly and even despair for a long, long time, and I finally had some expectations for the future. In the past year, my expectations have finally become concrete little by little, and I am so lucky and so happy to be at IC. I have met many like-minded friends and encountered lots of kind and lovely professors and supervisors. I want to especially thank Niamh, Corrina, and Philippa. Their support throughout the IRP process is invaluable. Even though I feel that the project I have accomplished might not make any impact in anything and it is just a milestone in my life, they continued to encourage me positively. I also owe my deepest gratitude to my parents, who have unconditionally supported my dreams. Lastly, many thanks to all my friends. "If there are 100 moments of disappointment in life, there will be 101 moments when friends come to save you." Words are powerless to express my gratitude, I can only simply say: Thank you very much. I genuinely love this place and every individual person I have met here. A saying that I once read comes to mind: "According to the ancient philosopher Shao Yong, everything in this world will be recreated exactly after 129,600 years." No matter whether this holds true or not, if I were allowed to phrase it romantically, I would say ' I am already looking forward to seeing you again.'

References

Barnett, T. P., Adam, J. C., & Lettenmaier, D. P. (2005). Potential impacts of a warming climate on water availability in snow-dominated regions. *Nature*, *438*(7066), Article 7066. https://doi.org/10.1038/nature04141

Capelli, A., Koch, F., Marty, C., Henkel, P., Schweizer, J. (2020). Snow water equivalent measurements with low-cost GNSS receivers along a steep elevation gradient in the East-ern Swiss Alps. EnviDat. https://www.doi.org/10.16904/envidat.186.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J-N. (2023): ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), DOI: 10.24381/cds.adbb2d47 (Accessed on DD-MMM-YYYY)

Copernicus Climate Change Service. (2019). *ERA5-Land hourly data from 2001 to present* [dataset]. ECMWF. https://doi.org/10.24381/CDS.E2161BAC

Dahiru, T. (2008). P – VALUE, A TRUE TEST OF STATISTICAL SIGNIFICANCE? A CAUTIONARY NOTE. *Annals of Ibadan Postgraduate Medicine*, *6*(1), 21–26.

Data—AQUARIUS WebPortal. (n.d.). Retrieved 31 August 2023, from https://bcmoe-prod.aquaticinformatics.net/

Jonas, T., Marty, C., & Magnusson, J. (2009). Estimating the snow water equivalent from snow depth measurements in the Swiss Alps. *Journal of Hydrology*, *378*(1), 161–167. https://doi.org/10.1016/j.jhydrol.2009.09.021

Kart | Sildre. (n.d.). Retrieved 31 August 2023, from https://sildre.nve.no/map?x=333918&y=7317662&zoom=6

Koch, F., Henkel, P., Appel, F., Mauser, W., Schweizer, J. (2018). GPS-derived data of SWE, HS and LWC and corresponding validation data. EnviDat. https://www.doi.org/10.16904/envidat.56.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., Reblitz-Richardson, O. 2020. Captum: A Unified and

Generic Model Interpretability Library for PyTorch. arXiv:2009.07896 [cs.LG]. Accessed 24 August 2023. https://captum.ai/.

Larson, K. M. and E. E. Small. (n.d.). *Daily Snow Depth and SWE from GPS Signal-to-Noise Ratios, Version 1* [dataset]. Retrieved 7 August 2023, from https://nsidc.org/data/nsidc-0722/versions/1

Liston, G. E. and M. Sturm. Global Seasonal-Snow Classification, Version 1. 2021, Distributed by National Snow and Ice Data Center. https://doi.org/10.5067/99FTCYYYLAQ0. Date Accessed 08-07-2023.

Marks, D., Domingo, J., Susong, D., Link, T., & Garen, D. (1999). A spatially distributed energy balance snowmelt model for application in mountain basins. *Hydrological Processes*, *13*(12–13), 1935–1959. https://doi.org/10.1002/(SICI)1099-1085(199909)13:12/13<1935::AID-HYP868>3.0.CO;2-C

McCreight, J. L., & Small, E. E. (2014). Modeling bulk density and snow water equivalent using daily snow depth observations. *The Cryosphere*, *8*(2), 521–536. https://doi.org/10.5194/tc-8-521-2014

Ntokas, K. F. F., Odry, J., Boucher, M.-A., & Garnaud, C. (2021). Investigating ANN architectures and training to estimate snow water equivalent from snow depth. *Hydrology and Earth System Sciences*, *25*(6), 3017–3040. https://doi.org/10.5194/hess-25-3017-2021

Organization, W. M. (2017). *Guide to Meteorological Instruments and Methods of Observation: (CIMO guide). 2014 edition, updated in 2017.[SUPERSEDED]* [Report]. World Meteorological Organization. https://doi.org/10.25607/OBP-432

Siirila-Woodburn, E. R., Rhoades, A. M., Hatchett, B. J., Huning, L. S., Szinai, J., Tague, C., Nico, P. S., Feldman, D. R., Jones, A. D., Collins, W. D., & Kaatz, L. (2021). A low-to-no snow future and its impacts on water resources in the western United States. *Nature Reviews Earth & Environment*, *2*(11), Article 11. https://doi.org/10.1038/s43017-021-00219-y

Sturm, M., Holmgren, J., & Liston, G. E. (1995). A Seasonal Snow Cover Classification System for Local to Global Applications. *Journal of Climate*, *8*(5), 1261–1283. https://doi.org/10.1175/1520-0442(1995)008<1261:ASSCCS>2.0.CO;2

Sturm, M., Taras, B., Liston, G. E., Derksen, C., Jonas, T., & Lea, J. (2010). Estimating Snow Water Equivalent Using Snow Depth Data and Climate Classes. *Journal of Hydrometeorology*, *11*(6), 1380–1394. https://doi.org/10.1175/2010JHM1202.1

Winkler, M., Schellander, H., & Gruber, S. (2021). Snow water equivalents exclusively from snow depths and their temporal changes: The Δsnow model. *Hydrology and Earth System Sciences*, *25*(3), 1165–1187. https://doi.org/10.5194/hess-25-1165-2021

Appendices

Appendix 1

The following table presents the detailed description of the meteorological variables.

Table 1. Description of meteorological variables

Names	**Units**	**Description**
Temperature	°C	The air temperature at 2m above the surface of land, sea or in-land waters.
Solar radiation	(J m^{-2})	The amount of solar radiation that reaches Earth's surface minus the amount that is reflected by the Earth's surface.
Precipitation	m	The total amount of accumulated liquid and frozen water falling to the Earth, including snow and rain.
Snowfall	m of water equivalent	The total amount of snow that has accumulated on Earth's surface.

Appendix 2

The *Figure 1* below shows the changes of data distribution after removing zero values selectively in the dataset.

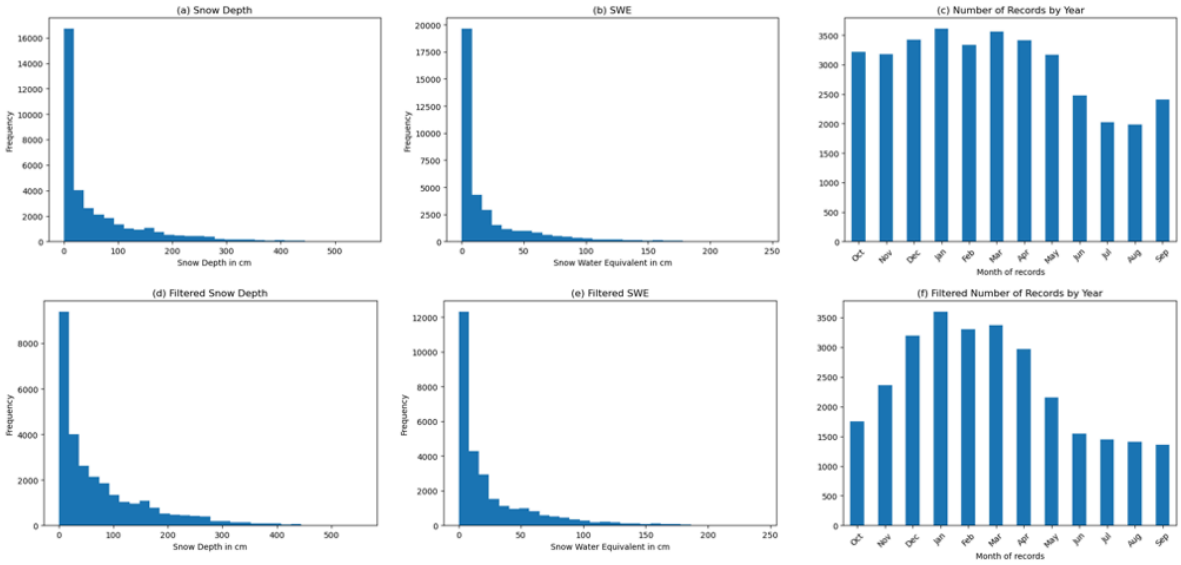

Figure 1. Distribution of the records in the training data for (a) Snow Depth (b) Snow Water Equivalent (c) The date of the records; the year starts on 1 October to cover the Norway, Switzerland, Canada and USA over years. (d) Snow Depth after removing zero value (e) Snow Water Equivalent after removing zero value (f) Number of Records after removing zero value

Appendix 3

The table below shows the detailed setup for the SLSTM and MLSTM models:

Table 1. The final setup for each model.

Characteristic	**SLSTM**	**MLSTM for different snow classes**				
		Tundra	Boreal Forest	Maritime	Prairie	Montane Forest
Number of hidden layers	2	3	2	3	2	1
Number of hidden neurons	[60, 30]	[70, 50, 30]	[60, 30]	[70, 50, 30]	[60, 30]	[50]
Number of epochs	60	100	30	200	50	100
Optimiser	Adam	Adam	Adam	Adam	Adam	Adam
Activation function	ReLU	ReLU	ReLU	ReLU	ReLU	ReLU
Loss function	MSE	MSE	MSE	MSE	MSE	MSE
Feature selection	Snow depth	Snow depth	Snow depth	Snow depth, Temperature, Precipitation	Snow depth	Snow depth
Time sequence (day)	30	14	10	30	10	30

Appendix 4

The following figures presents the detailed performance metrics results of various models' architectures.

Tundra

In *Figure 1*, it can be observed that the model with 3 layers and hidden neurons of 70, 50, and 30 respectively has the greater performances on RMSE, MBE, and nearly top value of R^2. The slightly higher MAE value potentially indicates that the model may predict better when estimating the large values, while there may have more error when dealing with small values than the other models. The model with 3 layers is selected because its overall predictive ability is still better than any other models as shown in *Figure 1*.

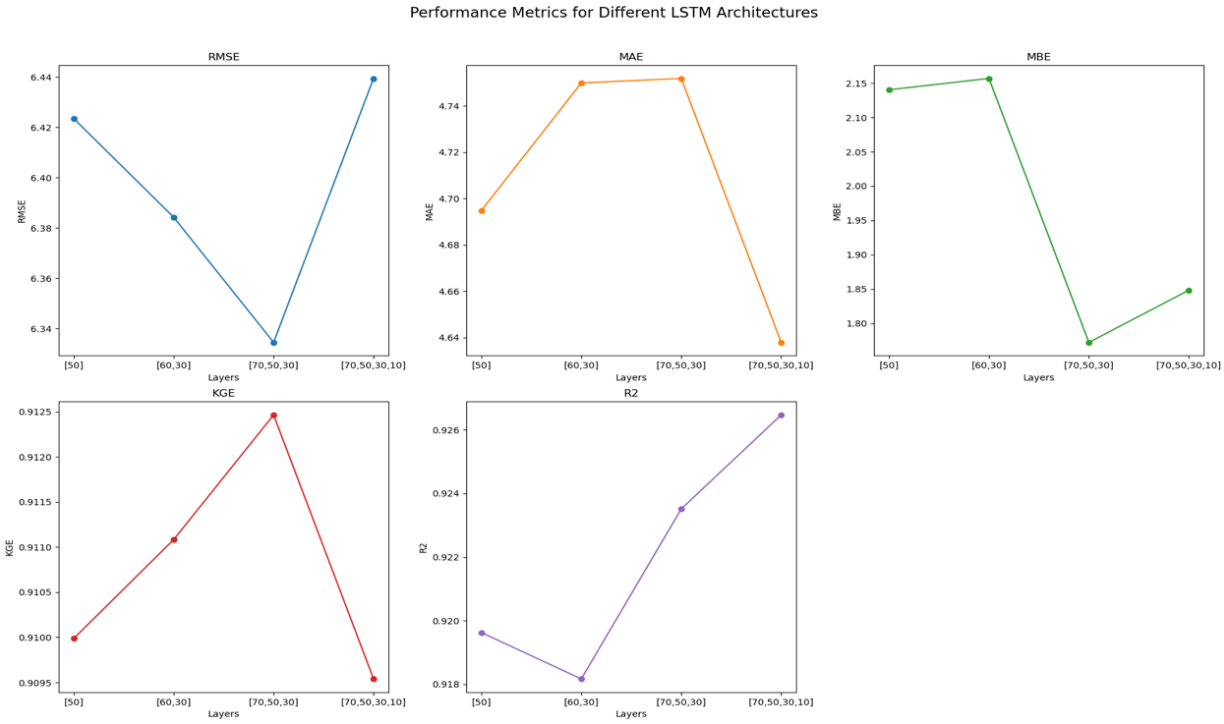

Figure 1. The performance metrics for different architectures for Tundra

Boreal Forest and Prairie

In *Figure 2*, it can be observed that both models with 2 layers and hidden neurons of 60, and 30 respectively has the greater performances on RMSE, MAE, and R^2. The low KGE values suggest that the models may not capture the dynamics changes of the data. For example, the observed data may have significant fluctuations, but the predicted values may be relatively stable. Both models with 2 layers are still selected because their overall predictive abilities are still better than any other models as shown in *Figure 2*.

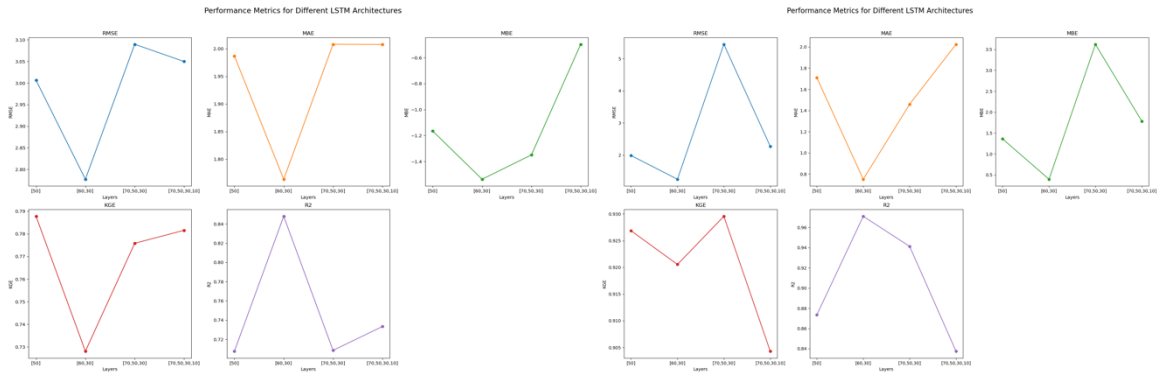

Figure 2. The performance metrics for different architectures for (a) Boreal Forest (b) Prairie

Montane Forest

In *Figure 3*, it can be observed that the model with only one layer and hidden neurons of 50 has the greater performance on RMSE, MAE, MBE, and R^2. While the KGE score is relatively low, the absolute value is close to the others. The model with one layer is selected because its overall predictive ability is better than any other models as shown in *Figure 3*.

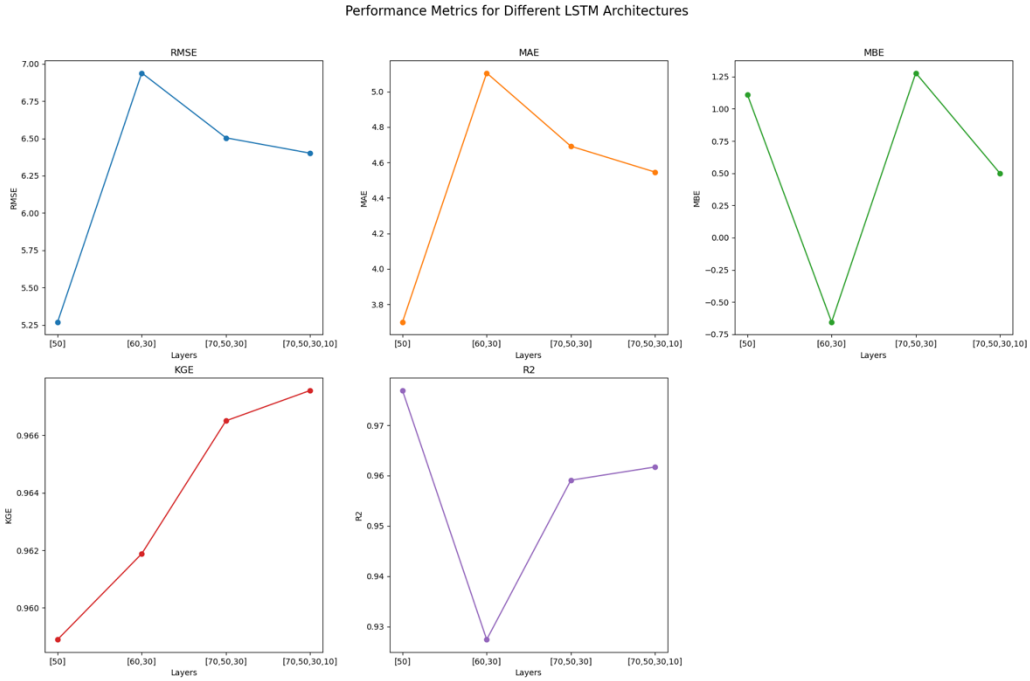

Figure 3. The performance metrics for different architectures for Montane Forest

Appendix 5

Tundra

The step is similar for snow class Tundra as above (for Maritime). In *Figure 2(a), Solar radiation* and *Precipitation* were almost the least significant features, hence we removed them from the subsequent training. Then we repeated this step iteratively. The results of the sensitivity analysis are presented in *Table 1(a)*. The model with *Snow depth* only shows the optimal result. *Snow depth* has the highest correlation as shown in Table 1(b), confirms the result. In addition, the experiments of time sequence were conducted. 7-day, 10-day, and 14-day were experimented as the attributions start to vary from around day 80 as shown in *Figure 2(b).* The time sequence is decided to be 14-day according to the performances shown in *Figure 1 of Appendix 6*.

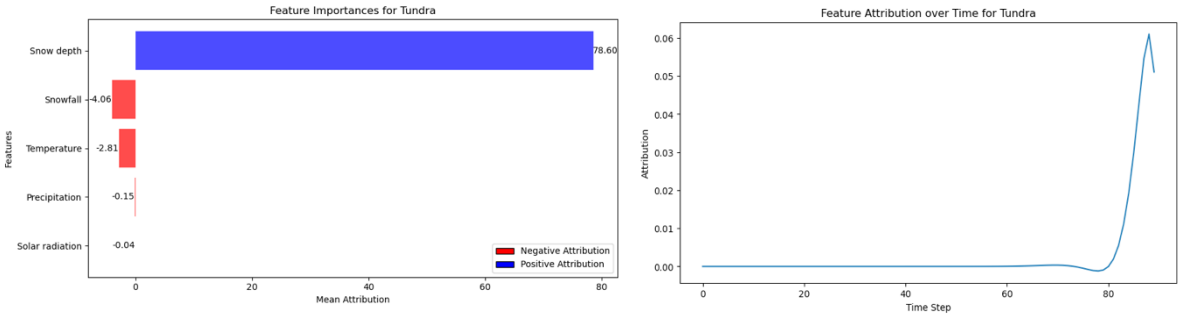

Figure 2. (a) Feature Attributes for Tundra over all samples and time steps, (b) Snow depth Attribute over 90 days

Table 1. (a) Performance evaluation metrics of MLSTM models for Tundra with different meteorological features (b) Pearson and Spearman correlation between all meteorological features and SWE for Tundra

Meteorological Features	**RMSE**	**MAE**	**MBE**	**KEG**	**R^2**
Snow depth, Temperature, Snowfall	5.830	4.470	0.505	0.927	0.927
Snow depth, Temperature	8.140	5.638	3.729	0.852	0.859
Snow depth	5.426	3.595	1.448	0.930	0.942

Variables	**Pearson correlation to SWE**	**Spearman correlation to SWE**
Snow depth	0.961	0.968
Precipitation	-0.020	-0.0160
Snowfall	0.099	0.309
Solar radiation	-0.081	-0.165
Temperature	0.313	-0.458

Boreal Forest

For the snow class Boreal Forest, the following conclusions can be drawn. *Figure 3(a)* presents that *Snowfall* and *Precipitation* were almost the least significant features, hence we removed them from the subsequent training. Then we repeated this step iteratively. The results of the sensitivity analysis are presented in *Table 3(a)*. The final model with *Snow depth* only demonstrates the optimal result. *Snow depth* has the highest correlation as shown in *Table 2(b),* confirms the result. In addition, the experiments of time sequence were conducted. 10-day, 14-day, and 30-day were experimented as the attributions start to vary from around day 70 as shown in *Figure 3(b).* The time sequence is decided to be 10-day according to the performances shown in *Appendix 6 Figure 2*.

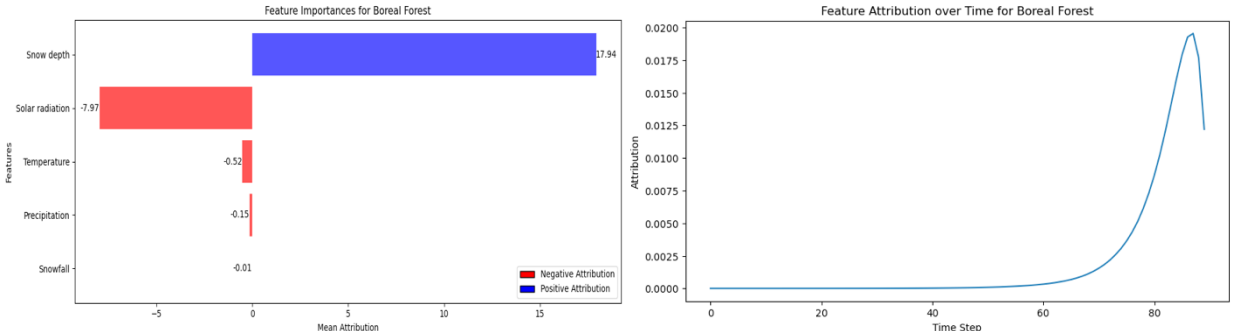

Figure 3. (a) Feature Attributes for Boreal Forest over all samples and time steps, (b) Snow depth Attribute over 90 days

Table 2. (a) Performance evaluation metrics of MLSTM models for Boreal Forest with different meteorological features (b) Pearson and Spearman correlation between all meteorological features and SWE for Boreal Forest

Meteorological Features	**RMSE**	**MAE**	**MBE**	**KEG**	R^2
Snow depth, Temperature, Solar radiation	4.965	3.940	2.972	0.661	0.409
Snow depth, Solar radiation	3.415	5.638	-1.015	0.709	0.720
Snow depth	2.290	1.303	-0.783	0.779	0.900

Variables	**Pearson correlation to SWE**	**Spearman correlation to SWE**
Snow depth	0.959	0.975
Precipitation	-0.018	-0.007
Snowfall	0.008	0.148
Solar radiation	-0.114	-0.053
Temperature	-0.0488	-0.243

Prairie

For the result of Prairie shown in *Figure 4(a),* it presents that *Snowfall* and *Precipitation* were almost the least significant features, hence we removed them from the subsequent training. Then we repeated this step iteratively. The results of the sensitivity analysis are presented in *Table 3(a)*. The final model with *Snow depth* only shows the optimal result. *Snow depth* has the highest correlation as shown in *Table 3(b)*, which confirms the result. In addition, the experiments of time sequence were conducted. 7-day, 10-day, and 14-day were experimented as the attributions start to vary from around day 80 as shown in *Figure 4(b).* The time sequence is decided to be 10-day according to the performances shown in *Appendix 6 Figure 3*.

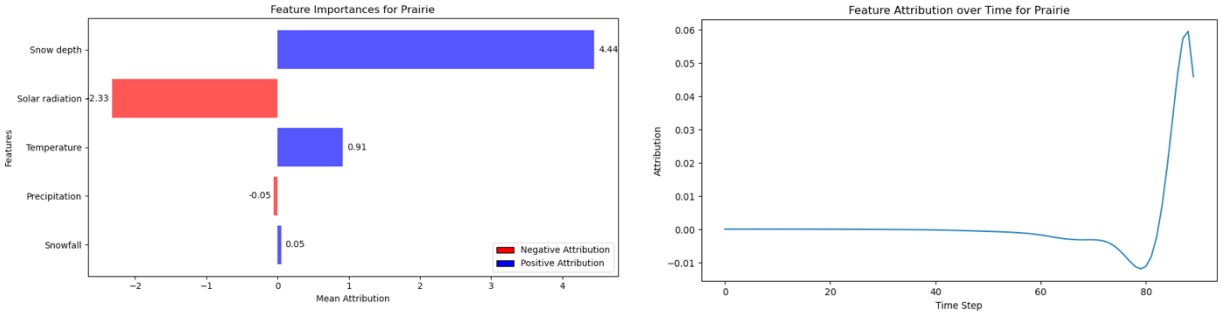

Figure 4. (a) Feature Attributes for Prairie over all samples and time steps, (b) Snow depth Attribute over 90 days

Table 3. (a) Performance evaluation metrics of MLSTM models for Prairie with different meteorological features (b) Pearson and Spearman correlation between all meteorological features and SWE for Prairie

Meteorological Features	**RMSE**	**MAE**	**MBE**	**KEG**	**R^2**
Snow depth, Temperature, Solar radiation	3.269	2.457	-2.192	0.763	0.800
Snow depth, Solar radiation	2.061	1.689	0.588	0.802	0.920
Snow depth	1.247	0.752	0.395	0.921	0.970

Variables	**Pearson correlation to SWE**	**Spearman correlation to SWE**
Snow depth	0.957	0.985
Precipitation	-0.005	0.033
Snowfall	0.027	0.180
Solar radiation	-0.032	-0.155
Temperature	-0.175	-0.298

Montane Forest

For the result of Montane Forest shown in *Figure 5(a),* it presents that *Snowfall* was the least significant features, hence we removed it from the subsequent training, and repeated the step iteratively. The results of the sensitivity analysis are shown in *Table 4(a)*. The final model with *Snow depth* only has the best performance. *Snow depth* also has the highest correlation as shown in *Table 3(b)*, which confirms the result. In addition, the experiments of time sequence were conducted. 7-day, 14-day, and 30-day were experimented as the attributions start to vary from around day 70 as shown in *Figure 3(b).* The time sequence is decided to be 30-day according to the performances shown in *Appendix 6 Figure 4*.

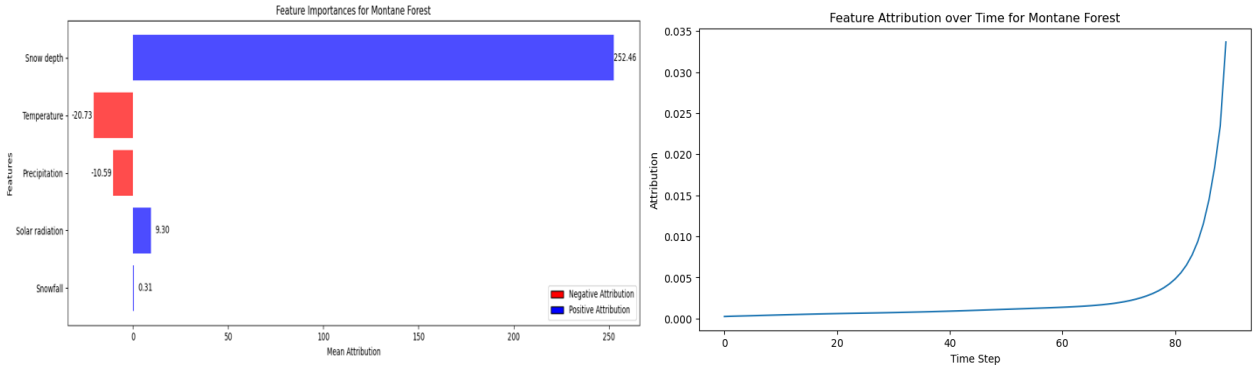

Figure 5. (a) Feature Attributes for Montane Forest over all samples and time steps, (b) Snow depth Attribute over 90 days

Table 4. (a) Performance evaluation metrics of MLSTM models for Montane Forest with different meteorological features (b) Pearson and Spearman correlation between all meteorological features and SWE for Montane Forest

Meteorological Features	**RMSE**	**MAE**	**MBE**	**KEG**	**R^2**
Snow depth, Temperature, Solar radiation	6.654	4.743	1.809	0.942	0.965
Snow depth, Precipitation	6.481	4.324	0.617	0.974	0.967
Snow depth	5.269	3.700	1.110	0.959	0.977

Variables	**Pearson correlation to SWE**	**Spearman correlation to SWE**
Snow depth	0.926	0.985
Precipitation	0.113	0.039
Snowfall	0.186	0.202
Solar radiation	-0.131	-0.204
Temperature	-0.347	-0.341

Appendix 6

Tundra

The grid search is utilised to methodologically evaluate different time sequences, including 7-day, 10-day, 14-day, and 20-day, the performance results are shown in *Figure 1*. Finally, 14-day was selected since it shows strong performance on RMSE, MAE, KGE and R^2. While MBE does not achieve the optimal result, its value remains within an acceptable range. The model with 14-day sequence demonstrates its accurate predictive ability.

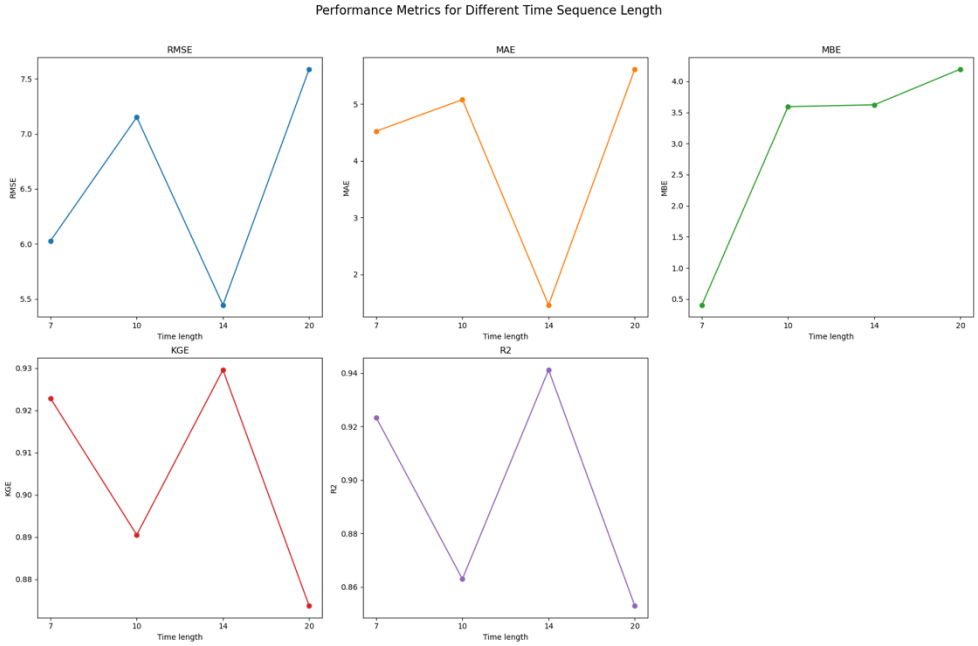

Figure 1. The performance metrics of models for different time sequence length for Tundra

Boreal Forest

The grid search is utilised to methodologically evaluate different time sequences, including 7-day, 10-day, 14-day, and 30-day, the performance results are shown in *Figure 2*. Finally, 10-day was selected since it shows strong performance on RMSE, MAE, KGE and R^2. While MBE does not achieve the

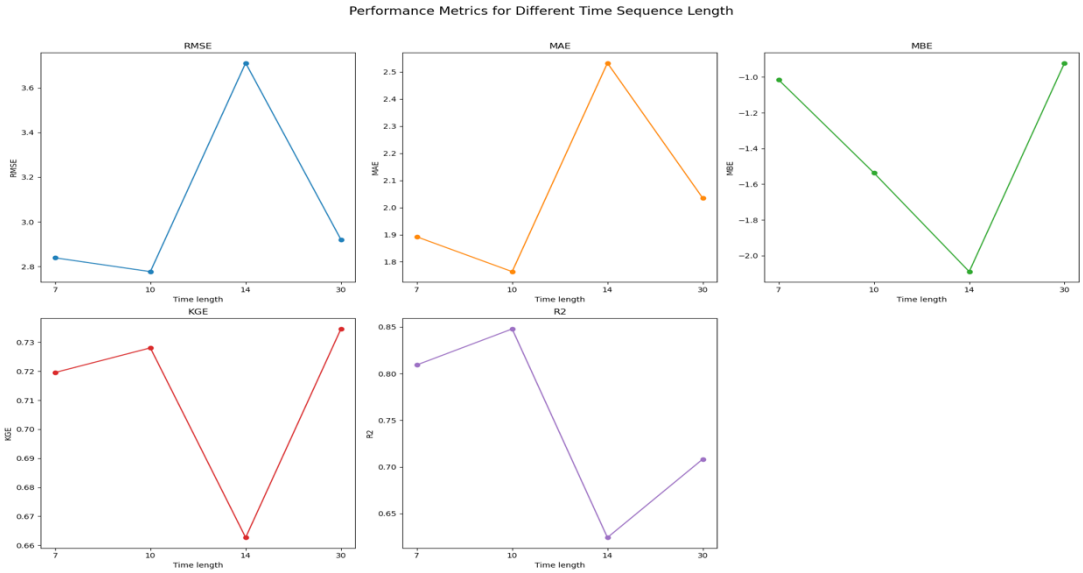

Figure 2. The performance metrics of models for different time sequence length for Boreal Forest

optimal result, its value remains within an acceptable range, showing an overall effectiveness.

Prairie

The grid search is utilised to methodologically evaluate different time sequences, including 7-day, 10-day, 14-day, and 20-day, the performance results are shown in *Figure 3*. Finally, 10-day model was selected since it shows great performance on RMSE, MAE, MBE and R^2. While KGE does not achieve the optimal result, its value remains within an acceptable range and is only marginally 0.02 different from 7-day model, showing an overall outstanding performance.

Montane Forest

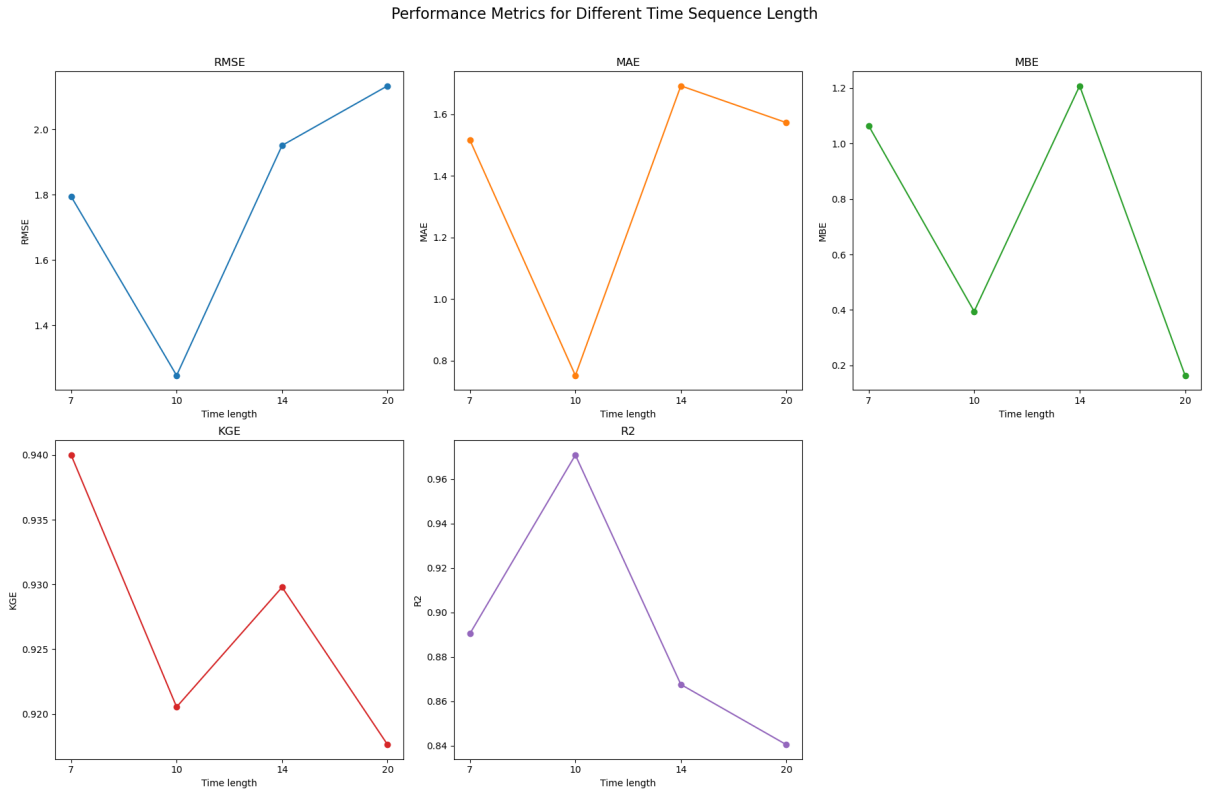

Figure 3. The performance metrics of models for different time sequence length for Prairie

The grid search is utilised to methodologically evaluate different time sequences, including 7-day, 14-day, 30-day, and 50-day, the performance results are shown in *Figure 4*. Finally, 30-day model was selected since it shows great performance on RMSE, MAE, KGE and R^2. While MBE does not achieve the optimal result, its value remains within an acceptable range which is second-place ranking. The model with 30-day sequence still has an overall outstanding performance.

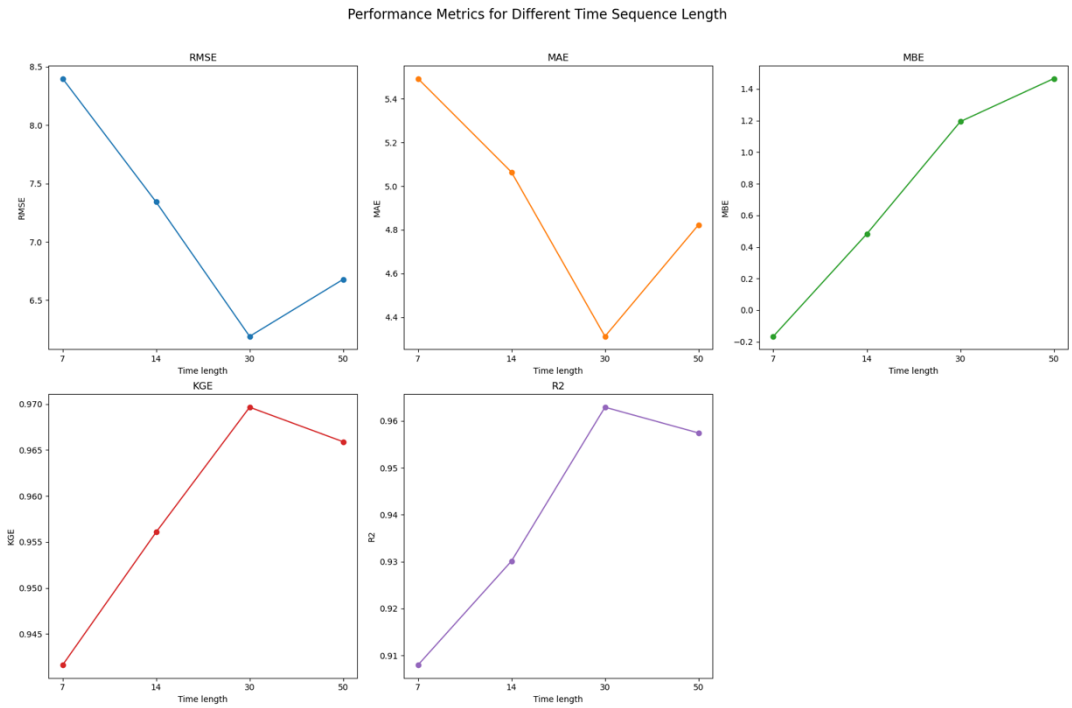

Figure 4. The performance metrics of models for different time sequence length for Montane Forest