BENCHMARKING PROBABILISTIC MACHINE LEARNING MODELS FOR ARCTIC SEA ICE FORECASTING

Sahara Ali, Seraj Al Mahmud Mostafa, Xingyan Li, Sara Khanjani, Jianwu Wang, James Foulds, Vandana Janeja

University of Maryland Baltimore County, MD 21250, USA

ABSTRACT

The Arctic is a region with unique climate features, motivating new AI methodologies to study it. Unfortunately, Arctic sea ice has seen a continuous decline since 1979. This not only poses a significant threat to Arctic wildlife and surrounding coastal communities but is also adversely affecting the global climate patterns. To study the potential of AI in tackling climate change, we analyze the performance of four probabilistic machine learning methods in forecasting sea-ice extent for lead times of up to 6 months, further comparing them with traditional machine learning methods. Our comparative analysis shows that Gaussian Process Regression is a good fit to predict sea-ice extent for longer lead times with lowest RMSE error.

***Index Terms*—** Arctic sea ice, climate change, probabilistic machine learning, Gaussian Process Regression

1. INTRODUCTION

The enormous areas of Arctic ice and snow are responsible for reflecting sunlight back to space which keeps the planet cool and regulates global and regional weather patterns [11]. However, the Arctic sea ice has seen a continuous decline since 1979 and is half of what it was in 1970. It is important to predict fluctuations in the Arctic sea ice by modeling the weather patterns as it can improve our understanding of potential changes facing the global climate. Since the climate data presents high spatiotemporal correlations, machine learning models have shown promising results in spatiotemporal data mining leading to short and long term weather forecasting [10]. In order to convince physicists of the potential of these data-driven approaches, however, quantifying model uncertainty still remains a major challenge. Probabilistic machine learning models, particularly Bayesian models, provide a principled approach for quantifying uncertainty. Though recent data-driven approaches have shown promising results in sea-ice forecasting, they still struggle with sub-seasonal forecasting at longer lead times [2, 3]. We therefore present a rigorous study of four probabilistic models and two baseline models to forecast sea ice extent at multiple lead times, further proposing directions to quantify epistemic uncertainty in model predictions. Our results illustrate the value of the probabilistic modeling approach in this context. We have open-sourced our work at github.[1] The contributions of our work include: *(i)* Customizing probabilistic machine learning models to forecast sea ice extent (SIE) at lead times of 1 to 6 months. *(ii)* Performing a rigorous study of the accuracy versus lead-time tradeoff for the implemented models. *(iii)* Benchmarking the performance of probabilistic and standard models to assist researchers in forecasting sea ice variations.

This work is partially supported by grants OAC–1942714 and OAC–2118285 from National Science Foundation.

2. METHODOLOGY

In this section, we will first explain the problem definition and then provide details on the dataset used and the methods implemented. The end-to-end pipeline of the benchmarking process is illustrated in Figure 1. A high-level summary of our comparative analysis is given in Table 1.

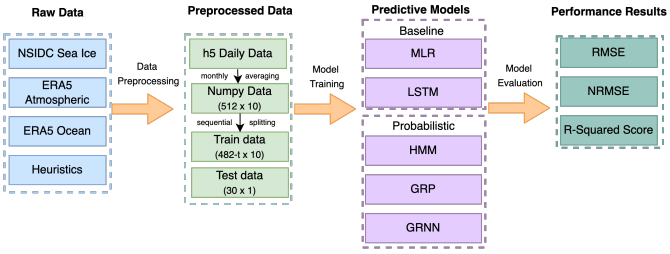

Fig. 1. End-to-end pipeline of our benchmarking experiments.

2.1. Problem Definition

Owing to the fluctuations in sea-ice over summer and winter seasons, accurate sub-seasonal forecasting becomes a great challenge. To tackle this challenge, we compare the predictions of all six models at a lead time of 1 to 6 months. We specifically look at the following problem definition: *Given **N** months of historic meteorological and sea-ice data X, learn a probabilistic function to forecast sea-ice extent Y for the next **M** months in future.*

$$Y_{t+m} = f(X_{t-n}, X_{t-n+1}, \ldots, X_t) \tag{1}$$

[1] https://github.com/big-data-lab-umbc/sea-ice-prediction

Table 1. Comparison of predictive models and their important characteristics

Methods	Purpose	Variables	Type	Probabilistic	Uncertainty Quantification
MLR	Regression technique; uses more than one independent variable to predict a response [1].	Multivariate	ML	No	Indirectly via model ensembling
LSTM	Sequence prediction or time-series forecasting; predicts future values over large time periods from sequence of data [5].	Multivariate	DL	No	Indirectly via dropout layers or model ensembling
GRNN	Regression technique; non-parametric method that provides estimates of continuous variable and function approximations [9].	Multivariate	DL	Yes	Yes
GPR	Regression technique; computes predictive distribution using Bayesian approach [8].	Multivariate	ML	Yes	Yes
HMM	Sequence prediction; capable of observing unknown facts using Markov process that contains hidden parameters [4].	Univariate	ML	Yes	Indirectly via model ensembling
BLR	Regression technique; outcome is drawn from a probability distribution instead of point estimates [7].	Multivariate	ML	Yes	Yes

2.2. Dataset

This study utilizes 512 temporal records of observational data for 42 years from 1979-2021 over the Arctic region. These records include monthly mean values of sea-ice extent (SIE) from Nimbus-7 SSMR and DMSP SSM/I-SSMIS passive microwave data version[2] and 9 meteorological data variables obtained from ERA-5 global reanalysis product.[3] The choice and details of these variables is presented in our previous causal discovery study conducted on Arctic sea-ice [6]. To conduct our experiments, we first combined all the raw variable datasets to have single temporal and spatial resolution, i.e. monthly means and 1 degree (180 lat.$\times$360 lon.) of spatial resolution. Next we replaced missing values with interpolated values for sea-ice and replaced other missing values with 0. Finally we normalized data using the MinMax normalization technique.

2.3. Baseline Models

We implemented two widely used predictive models as baseline methods to compare with our probabilistic models. These include the Multiple Linear Regression (MLR) and Long Short Term Memory Model (LSTM).

2.3.1. Multiple Linear Regression (MLR)

We trained a Multiple Linear Regression model to predict sea ice extent values with a sequential split on the data containing all 10 variables including sea ice extent as input features. The predicted sea ice extent values are for a lead time of M months. Here, M is from 1 to 6. Six MLR models are trained, one for each lead time.

2.3.2. Long Short Term Memory Model (LSTM)

LSTM is a variant of the Recurrent Neural Network (RNN) used for time series data analysis and forecasting. Our LSTM based network comprises two many-to-many LSTM and one many-to-one LSTM layer, one dropout layer and three fully-connected layers. Like MLR, we trained six LSTM models for 6 months' lead times and optimized them using the 'Adam' optimizer.

2.4. Probabilistic Models

We implemented following four probabilistic models for forecasting sea ice extent for lead times of 1 to 6 months.

2.4.1. General Regression Neural Network (GRNN)

GRNN is a modified form of a radial basis network (RBF), that estimates values for continuous variables using nonparametric estimators of probability density functions (PDF) [9]. GRNN uses an extra layer of summation and weight connection between the hidden and output layer. The summation layer takes the input from previous layers and sums those values together to form the probability density function (PDF) using the Parzen window. GRNN uses Equation 2 where y is the estimator output, x is the estimator input vector and $E[y|x]$ is the expected value of output. For the given input vector x, $f(x, y)$ would be the joint PDF of x and y:

[2] NSIDC (nsidc.org/data/NSIDC-0051)
[3] ECMWF (cds.climate.copernicus.eu/cdsapp!/home)

$$E[y|x] = \frac{\int_{-\infty}^{\infty} y.f(x,y))dy}{\int_{-\infty}^{\infty} f(x,y))dy}. \quad (2)$$

2.4.2. *Hidden Markov Model (HMM)*

HMM is a statistical model able to observe hidden events using a Markov process. It consists of *(i)* a sequence of observable variables, *(ii)* a sequence of hidden states, *(iii)* a transition matrix, and *(iv)* an emission matrix to explain distribution of observed variables generated from the hidden states. The transition matrix is composed of probability for each hidden state transforming to the other, while the emission matrix comprises probabilities for each observed variable to be generated by the corresponding hidden states. We implemented the GaussianHMM variant of the model as it handles continuous distributions for observation state sequence. The number of hidden states is kept at six as the sea ice extent of six lead times need to be estimated, and the fractional changes are taken as features.

2.4.3. *Gaussian Process Regression (GPR)*

We find GPR a good fit for our problem since it is a popular non-parametric probabilistic model for regression with a Bayesian approach. Instead of calculating the probability distribution of parameters of a specific function, GPR calculates the probability distribution over all admissible functions that fit the data. Therefore, when we fit GPR on our training data it estimates the posterior distribution. Given features X and outcome y, GPR calculates the posterior distribution using Bayes' rule as given in Equation 3 where w represents the model parameters learnt. While fitting GPR, the log-marginal-likelihood (LML) is optimized which helps learn the parameters of the kernel. We implemented GPR by setting the prior mean of the GP to be the training data's mean. We repeated the training procedure 10 times to prevent LML from being trapped in local optima.

$$p(w|y,X) = \frac{p(y|X,w)p(w)}{p(y|X)} \quad (3)$$

2.4.4. *Bayesian Linear Regression (BLR)*

Lastly, we implemented BLR to test its usability on the problem domain. In contrast to MLR, here both the output y and the model parameters are generated from a normal (Gaussian) distribution characterized by a mean and variance. First, we specified priors for the model parameters using normal distribution, then we created a model mapping the training inputs to the training outputs, finally we used a Markov Chain Monte Carlo (MCMC) algorithm to draw 500 samples from the posterior distribution for the model parameters. The end result is an estimate of the posterior distribution for the parameters using which we can compute predicted outcomes for a given test dataset.

2.5. Uncertainty Quantification

A valuable feature of probabilistic models is their ability to predict probability distributions instead of discrete values. We can estimate model uncertainty by computing variance and standard deviation of the posterior predictive probabilities given by the models. In case of non-probabilistic models, the epistemic uncertainty can be quantified by ensembling predictions from multiple copies of the same model.

3. RESULTS AND DISCUSSION

We trained all models on the first 40 years of data and tested them on last 30 months, that is, from February 2019 to August 2021. We evaluated their performance by calculating the Root Mean Square Error (RMSE) (in million square kilometers). To tackle such large values, we normalized RMSE scores by dividing RMSE values with the mean of observed sea ice extent, that is, $\bar{y}$. Lower the values of RMSE and NRMSE, better the predictive performance. We further investigated the variations in predicted SIE values due to input features using the correlation co-efficient score, also known as R-Squared or R2 Score. Higher the values of R2 Score, better the performance.

Looking at the results in Table 2, it is evident that the prediction error increases with the increase in lead times. The increasing RMSE implies that models' performance are greatly affected by seasonal patterns, however, we noticed different patterns in the performance of different models.

Evaluating the performance of deep learning models, we see that GRNN outperformed LSTM – one of the widely used methods for Arctic sea-ice prediction [2]. This implies that GRNN can be a good alternative for learning non-linearities in data alongwith uncertainty prediction over longer lead times. In case of HMM, as it follows the forward-backward algorithm, the steady climbing in the RMSE scores makes sense because the prediction of each observation strongly depends on the previous states; also evident by negative R2 Score. Further, we observed that GPR gets better from one lead time to another, which makes it the best fit for the ice extent prediction for different lead times. This commendable performance could be due to GPR's underlying Bayesian approach, which infers a probability distribution over all possible values, irrespective of the sequential patterns. However, BLR performed against our expectations and had the highest error for all lead times, making it unfit for the problem domain. We did not investigate BLR's failure to reach convergence but it can be inspected further in future studies. Qualitatively, we observed in Figure 2 that both GPR and GRNN better captured the peak values during Summer and Winter seasons, however GPR outperformed GRNN for the lead time of 6 months.

4. CONCLUSION AND FUTURE WORK

In this paper, we compared six of the renowned probabilistic and standard machine learning models suitable for forecast-

Table 2. RMSE, NRMSE and R-Squared scores for all models for lead times of 1 to 6 months.

	RMSE						NRMSE						R-Squared					
Lead time (months)	**1**	**2**	**3**	**4**	**5**	**6**	**1**	**2**	**3**	**4**	**5**	**6**	**1**	**2**	**3**	**4**	**5**	**6**
MLR	**0.4331**	0.8928	1.4292	1.7148	2.0042	2.1663	**0.0412**	0.0850	0.1361	0.1633	0.1908	0.2062	**0.985**	0.937	0.839	0.769	0.684	0.631
HMM	1.6535	3.3911	4.7954	6.3511	7.0038	7.2644	0.1558	0.3201	0.4540	0.6038	0.6715	0.7018	0.783	0.081	-0.860	-2.286	-2.931	-3.066
GPR	0.4523	**0.4514**	**0.4055**	**0.3985**	**0.4007**	**0.3936**	0.0430	**0.0720**	**0.0978**	**0.1113**	**0.1197**	**0.1256**	0.983	**0.955**	**0.917**	**0.892**	**0.875**	**0.863**
LSTM	0.6274	1.1022	1.4313	1.5610	1.7222	1.7636	0.0697	0.1060	0.1338	0.1561	0.1517	0.2025	0.958	0.903	0.845	0.789	0.801	0.645
GRNN	0.5784	0.8048	1.0311	1.2550	1.5549	1.8162	0.0550	0.0766	0.0981	0.1195	0.1480	0.1729	0.973	0.949	0.916	0.876	0.810	0.740
BLR	8.0842	6.5638	8.4614	7.2429	8.3365	6.6301	0.7618	0.5307	0.7973	0.6825	0.7856	0.6248	-4.815	-2.834	-5.371	-3.668	-5.184	-2.911

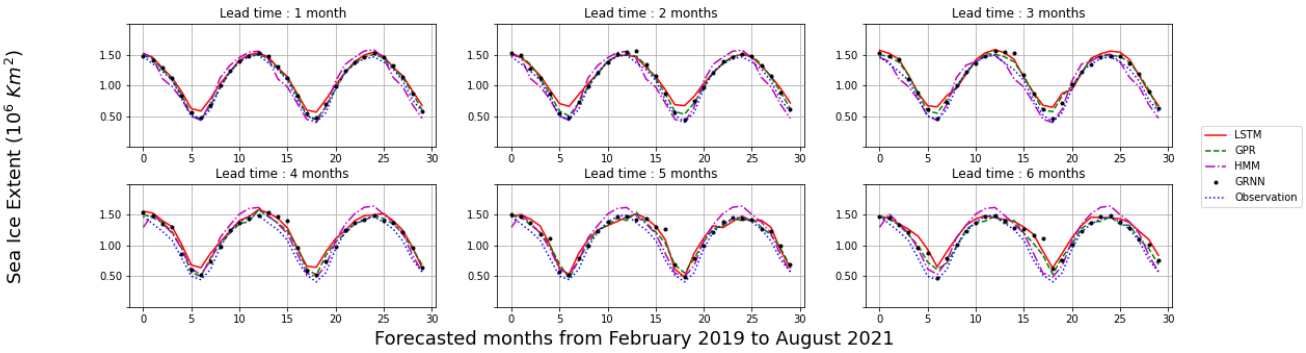

Fig. 2. Time-series predictions from five models versus the observations for lead time of 1 to 6 months.

ing sea-ice extent at greater lead times. We conclude that: *(i)* Probabilistic models can be a good alternative to deep learning models for small datasets. *(ii)* All models' performance decreases as lead times increase. However, GPR outperformed MLR and LSTM with a 12% increase in R2-Score making it the best fit for the ice extent prediction at greater lead times. *(iii)* HMM and BLR are not suitable for seasonal forecasting at greater lead times. In future, we plan to extend the benchmarking on spatiotemporal data incorporating higher dimensions in the performance evaluation.

5. REFERENCES

[1] L. S. Aiken, S. G. West, and S. C. Pitts. Multiple linear regression. *Handbook of Psychology*, pages 481–507, 2003.

[2] S. Ali, Y. Huang, X. Huang, and J. Wang. Sea Ice Forecasting using Attention-based Ensemble LSTM. *ArXiv e-prints*, page arXiv:2108.00853, July 2021.

[3] T. R. Andersson, J. S. Hosking, M. Pérez-Ortiz, B. Paige, A. Elliott, C. Russell, S. Law, D. C. Jones, J. Wilkinson, T. Phillips, et al. Seasonal Arctic sea ice forecasting with probabilistic deep learning. *Nature Communications*, 12(1):1–12, 2021.

[4] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563, 1966.

[5] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[6] Y. Huang, M. Kleindessner, A. Munishkin, D. Varshney, P. Guo, and J. Wang. Benchmarking of Data-Driven Causality Discovery Approaches in the Interactions of Arctic Sea Ice and Atmosphere. *Frontiers in Big Data*, 4, 2021.

[7] T. Minka. Bayesian linear regression. Technical report, Citeseer, 2000.

[8] J. Quinonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

[9] D. F. Specht et al. A general regression neural network. *IEEE Transactions on Neural Networks*, 2(6):568–576, 1991.

[10] S. Wang, J. Cao, and P. Yu. Deep Learning for Spatio-Temporal Data Mining: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2020.

[11] M. Wendisch, A. Macke, A. Ehrlich, C. Lüpkes, M. Mech, D. Chechin, K. Dethloff, C. B. Velasco, H. Bozem, M. Brückner, et al. The Arctic cloud puzzle. *Bulletin of the American Meteorological Society*, 100(5):841–871, 2019.