Sub-Seasonal Climate Forecasting via Machine Learning: Challenges, Analysis, and Advances

Sijie He [*1] **, Xinyan Li** [*1]**, Timothy DelSole** [2]**, Pradeep Ravikumar** [3]**, Arindam Banerjee** [4]

[1]Department of Computer Science & Engineering, University of Minnesota, Twin Cities
[2] Department of Atmospheric, Oceanic, and Earth Science, George Mason University
[3] Machine Learning Department, Carnegie Mellon University
[4] Department of Computer Science, University of Illinois Urbana-Champaign
{hexxx893, lixx1166}@umn.edu, tdelsole@gmu.edu, pradeepr@cs.cmu.edu, arindamb@illinois.edu

Abstract

Sub-seasonal forecasting (SSF) focuses on predicting key variables such as temperature and precipitation on the 2-week to 2-month time scale. Skillful SSF would have immense societal value in such areas as agricultural productivity, water resource management, and emergency planning for extreme weather events. However, SSF is considered more challenging than either weather prediction or even seasonal prediction, and is still a largely understudied problem. In this paper, we carefully investigate 10 Machine Learning (ML) approaches to sub-seasonal temperature forecasting over the contiguous U.S. on the SSF dataset we collect, including a variety of climate variables from the atmosphere, ocean, and land. Because of the complicated atmosphere-land-ocean couplings and the limited amount of good quality observational data, SSF imposes a great challenge for ML despite the recent advances in various domains. Our results indicate that suitable ML models, e.g., XGBoost, to some extent, capture the predictability on sub-seasonal time scales and can outperform the climatological baselines, while Deep Learning (DL) models barely manage to match the best results with carefully designed architecture. Besides, our analysis and exploration provide insights on important aspects to improve the quality of sub-seasonal forecasts, e.g., feature representation and model architecture. The SSF dataset and code [1] are released with this paper for use by the broader research community.

1 Introduction

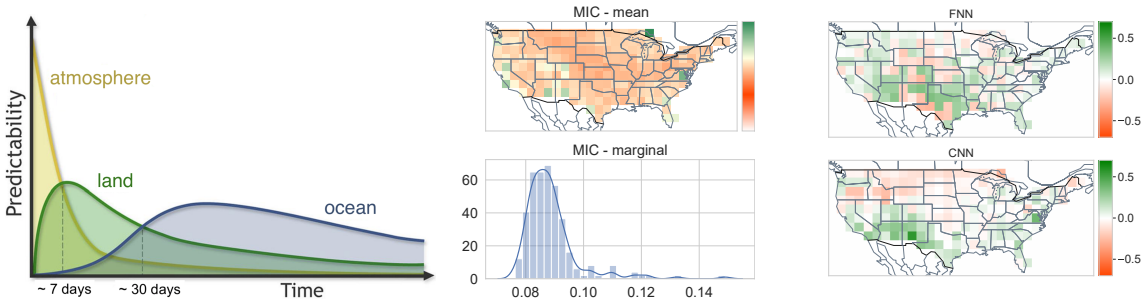
Over the past few decades, major advances have been made in weather forecasts on time scales of days to about a week (Lorenc 1986; Simmons and Hollingsworth 2002; National Research Council 2010). Similarly, major advances have been made in seasonal forecasts on time scales of 2-9 months (Barnston et al. 2012). However, making high-quality forecasts of key climate variables such as temperature and precipitation on sub-seasonal time scales, defined as the time range between 2-8 weeks, has long been a gap in operational forecasting (National Academies of Sciences 2016). Skillful climate forecasts at sub-seasonal time scales would be of immense societal value, and would have an impact in a wide variety of domains including agricultural productivity, water resource management, and emergency planning for extreme weather events, etc. (Pomeroy et al. 2002; Klemm and McPherson 2017). The importance of sub-seasonal climate forecasting (SSF) has been discussed in great detail in two recent high profile reports from the National Academy of Sciences (National Research Council 2010; National Academies of Sciences 2016). Despite the scientific, societal, and financial importance of SSF, progress on the problem has been limited (Braman et al. 2013; de Perez and Mason 2014), partly because it has attracted less attention compared to weather and seasonal climate prediction. Also, SSF is arguably more difficult compared to weather or seasonal forecasting due to limited predictive information from land and ocean, and virtually no predictability from the atmosphere (Uccellini and Jacobs 2018), which forms the basis of numerical weather prediction (Simmons and Hollingsworth 2002) (Figure 1(a)).

There exists great potential to advance sub-seasonal prediction using Machine Learning (ML) techniques. Due in large part to this potential promise, a recently concluded real-time forecasting competition called the Sub-Seasonal Climate Forecast Rodeo was sponsored by the Bureau of Reclamation in partnership with NOAA, USGS, and the U.S. Army Corps of Engineers (Raff et al. 2017; Hwang et al. 2019). However, a direct application of standard black-box ML approaches to SSF can run into challenges due to the high-dimensionality and strong spatial correlation of the raw data from atmosphere, ocean, and land, e.g., Figure 1(c) shows that popular approaches such as Fully connected Neural Networks (FNN) and Convolutional Neural Networks (CNN) do not perform so well when directly applied to the raw data. One reason is that sub-seasonal forecasting does not lie in the big data regime: about 40 years of reliable data exists for all climate variables, with each day corresponding to one data point, which totals less than 20,000 data points. Additionally, different seasons may have different predictive relations, and many climate variables have strong temporal correlations at daily time scales, further reducing the effective data size. Therefore, it is worth carefully and sys-

[*]The first two authors have equal contribution.

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]The SSF dataset and codebase are publicly available at https://sites.google.com/view/ssf-ml/home.



(a) Sources of Predictability (b) MIC (c) Results of FNN and CNN

Figure 1: (a) Sources of predictability at different forecast time scales. Atmosphere is most predictive on weather time scales, whereas for SSF, land and ocean are considered important sources of predictability (Uccellini and Jacobs 2018). (b) Maximum information coefficient (MIC) (Reshef et al. 2011) between residualized temperatures of week 3 & 4 and week -2 & -1. Small MICs ($\leq$ 0.1) at a majority of locations indicate little information shared between the most recent date and the forecasting target. (c) Predictive skills of Fully connected Neural Networks (FNN) and Convolutional Neural Networks (CNN), in terms of temporal cosine similarity (see definition in Section 5), for temperature prediction over 2017-2018. FNN and CNN do not perform well, as the cosine similarities for most locations are either negative (red) or close to zero (white).

tematically investigating the capability of ML approaches including Deep Learning (DL) while keeping in mind the high-dimensionality, spatial-temporal correlations, and limited observational data available for SSF. Our main contributions of this paper are as follows:

- We illustrate that, with the limited predictability at sub-seasonal time scale and the unique nature of climate data, i.e., strong spatial-temporal correlation, high-dimensionality, and limited amount of high-quality observational data, SSF imposes a great challenge for ML despite the recent advances in various domains.
- We perform a comprehensive empirical study on 10 ML approaches to SSF over the contiguous U.S. and show that suitable ML models, e.g., XGBoost, to some extent, capture predictability at sub-seasonal time scales and outperform existing approaches in climate science, such as climatology, i.e., the 30-year average at a given location and time. Notably, DL models are only able to match the best results after careful selection of architecture.
- We analyze and explore various aspects, e.g., feature representation and model architecture, which shed light on potential directions to improve the quality of sub-seasonal forecasts. An analysis of feature importance suggests that ocean and land covariates are more useful than atmospheric covariates, which is consistent with Figure 1(a).
- We construct an SSF dataset covering the contiguous U.S. and including climate variables from the atmosphere, ocean, and land. We release the dataset and a flexible code base for data extraction, preprocessing, and SSF model training and evaluation.

Organization of the paper. We discuss related work in Section 2. In Section 3, we describe the SSF problem tackled in this paper and demonstrate its difficulties. In Section 4, we outline the investigated ML approaches. The details of experimental setup and results are provided in Section 5 and Section 6, and we conclude in Section 7.

2 Related Work

Although statistical models were used for weather prediction before the 1970s (Frederik Nebeker 1995), since the 1980s weather forecasts mainly relied on physics-based dynamical models (Barnston et al. 2012). More recently, there has been a surge of applications for ML approaches to both short-term weather forecasting (Cofino et al. 2002; Grover, Kapoor, and Horvitz 2015; Radhika and Shashi 2009), and longer-term climate prediction (Badr, Zaitchik, and Guikema 2014; Cohen et al. 2019). However, little attention has been paid on forecasting on sub-seasonal time scale (Vitart, Robertson, and Anderson 2012). Recently, ML techniques have made great strides in statistical prediction in many fields, so it is natural to investigate whether it can advance sub-seasonal climate prediction. In particular, many advances have occurred in developing prediction models using spatiotemporal climate data (Steinhaeuser, Chawla, and Ganguly 2011; Goncalves, Banerjee, and Von Zuben 2017; Hwang et al. 2019), e.g., predicting land temperature using oceanic data (DelSole and Banerjee 2017; He et al. 2019).

Since SSF can be formulated as a sequential modeling problem (Sutskever, Vinyals, and Le 2014; Venugopalan et al. 2015), bringing the core strength of DL-based sequential modeling has great potential for a transformation in climate forecasting (Ham, Kim, and Luo 2019; Reichstein et al. 2019; Schneider et al. 2017). In the past decade, recurrent neural network (RNN) (Funahashi and Nakamura 1993) and long short-term memory (LSTM) models (Gers, Schmidhuber, and Cummins 2000) have become popular sequential models and have been successfully applied in language modeling and other seq-to-seq tasks (Sundermeyer, Schlüter, and Ney 2012). Starting from (Sutskever, Vinyals, and Le 2014; Srivastava, Mansimov, and Salakhudinov 2015), the encoder-decoder structure with RNN or LSTM has become one of the most competitive algorithms for sequence transduction. Variants of such models that incorporate mechanisms like convolution (Xingjian et al. 2015; Shi et al.

2017) or attention mechanisms (Bahdanau, Cho, and Bengio 2015) have achieved remarkable breakthroughs for audio synthesis, word-level language modeling, and machine translation (Vaswani et al. 2017).

SSF is an extremely important but understudied problem and ML is just starting to get used in this area. Within ML, Hwang et al. (2019) are the first to specifically focus on SSF over western U.S. and released their benchmark dataset. In this paper, we *expand* the spatial forecasting range to the entire contiguous U.S. and *extend* the set of predictors by including climate variables considered as important sources of predictability on sub-seasonal time scale (Uccellini and Jacobs 2018), such as soil moisture, Niño and NAO indices.

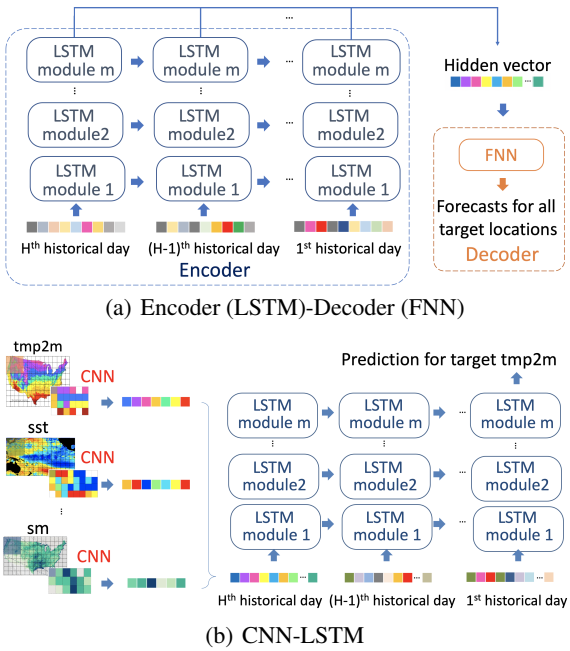
3 Sub-seasonal Climate Forecasting

Problem statement. In this paper, we focus on building temperature forecasting models at the forecast horizon of 15-28 days ahead, i.e., the target variable is the residualized average temperature of week 3 & 4. The geographic region of interest is the contiguous U.S. (latitudes 25N-49N and longitudes 76W-133W) at a 2° by 2° resolution (197 grid points). Our covariates consist of climate variables, such as sea surface temperature, soil moisture, geopotential height, etc., indicating the status of land, ocean, and atmosphere. Table 1 provides a detailed description.

Difficulty of the problem. To illustrate the challenge of SSF, we measure the statistical dependence between the residualized average temperature of week -2 & -1 (1-14 days in the past) and week 3 & 4 (15-28 days in the "future") at each grid point by maximum information coefficient (MIC) (Reshef et al. 2011), an information theory-based measure of the linear or non-linear association between two variables. The values of MIC range between 0 and 1, and a small MIC value close to 0 indicates a weak dependence. To assess statistical significance, we apply moving block bootstrap (Kunsch 1989) to time series of 2-week average temperature at each grid point from 1986 to 2018, with the block size of 365 days. The top panel in Figure 1(b) illustrates the average MIC from 100 bootstrap over the contiguous U.S., and the marginal distribution of all locations is shown at bottom. Small MIC values (≤ 0.1), indicating little predictive information shared between the most recent data and the forecasting target, to some extent, demonstrate how difficult SSF is.

From an ML perspective, applying black-box DL approaches naively to SSF is less likely to work due to the limited number of samples, and the high-dimensional and spatial-temporally correlated features. Figure 1(c) shows the performance of two vanilla DL models: FNN with ReLU activation function and CNN, in terms of the (temporal) cosine similarity between the prediction and the ground truth at each location over 2017-2018. For most locations, their cosine similarities are either negative or close to zero. Besides, as we illustrate in the sequel, we explore about 10 ML models for the problem, and most do not even get positive relative R^2, indicating they perform no better than the long term average (details are presented in Appendix A) [2]. Such results further demonstrate the difficulty of SSF.

[2] Appendix can be found at https://arxiv.org/abs/2006.07972.



(a) Encoder (LSTM)-Decoder (FNN)

(b) CNN-LSTM

Figure 2: Architectures of the designed DL models. (a) Encoder (LSTM)-Decoder (FNN) includes a few LSTM layers as the Encoder, and two fully connected layers as the Decoder. (b) CNN-LSTM consists of a few convolutional layers followed by an LSTM.

4 Methods

Notation. Let t denote a date and g denote a location. The target variable at time t and location g is the residualized average temperature over weeks 3 & 4 (from $t+15$ to $t+28$), denoted as $y_{g,t}$. For a given location g, $\mathbf{y}_{g,T}$ represents the target variable over a time range T. Similarly, $\mathbf{y}_{G,t}$ denotes the target variable over all G locations at time t. $X_t \in \mathbb{R}^p$ denotes the p-dimensional covariates at time t.

Non-DL models. We explore the following non-DL models.

- **MultiLLR** (Hwang et al. 2019). MultiLLR introduces a multitask feature selection algorithm to remove the irrelevant predictors and integrates the remaining predictors linearly. For a location g and a target date t^*, its coefficient β_g is estimated by $\hat{\beta}_g = argmin_\beta \sum_{t\in\mathcal{D}} w_{g,t}(y_{g,t} - \beta^T X_t)^2$, where $\mathcal{D}$ is the temporal span around the target date's day of the year and $w_{g,t}$ is the corresponding weight. In (Hwang et al. 2019), an equal data point weighting ($w_{g,t} = 1$) has been employed.
- **AutoKNN** (Hwang et al. 2019). An auto-regression model with weighted temporally local samples, where the auto-regression lags are selected via a multitask k-nearest neighbor criterion. The method only takes historical measurements of the target variables as input. The nearest neighbors of each target date are selected based on an average of spatial cosine similarity computed over a history of $M = 60$ days, starting one year prior to a target date t^* (lag $l = 365$). More pre-

Type	Climate variable	Description	Spatial coverage	Data Source
Spatiotemporal	tmp2m	Daily temperature at 2 meters	Contiguous U.S.	CPC Global Daily Temperature (Fan and Van den Dool 2008)
	sm	Monthly soil moisture		CPC Soil Moisture (Fan and van den Dool 2004)
	sst	Daily sea surface temperature	North Pacific & Atlantic Ocean	Optimum Interpolation SST (OISST) (Reynolds et al. 2007)
	rhum	Daily relative humidity	Contiguous U.S. and North Pacific & Atlantic Ocean	Atmospheric Research Reanalysi Dataset (Kalnay et al. 1996)
	slp	Daily pressure at sea level		
	hgt10 & hgt500	Daily geopotential height		
Temporal	MEI	Bimonthly multivariate ENSO index	NA	NOAA ESRL MEI.v2 (Zimmerman, Vimont, and Block 2016)
	Niño 1+2, 3, 3.4, 4	Weekly Oceanic Niño Index (ONI)		NOAA National Weather Service, CPC (Reynolds et al. 2007)
	NAO	Daily North Atlantic Oscillation index		NOAA National Weather Service, CPC (Van den Dool, Saha, and Johansson 2000)
	MJO phase & amplitude	Madden-Julian Oscillation index		Australian Government BoM (Wheeler and Hendon 2004)

Table 1: Description of climate variables and their data sources.

cisely, the similarity between the target date t^* and a date t in the corresponding training set is formulated as $\text{sim}_t = \frac{1}{M}\sum_{m=0}^{M-1}\cos(\mathbf{y}_{G,t-l-m}, \mathbf{y}_{G,t^*-l-m})$, where $\cos(\mathbf{y}_{G,t_1}, \mathbf{y}_{G,t_2})$ computes the (spatial) cosine similarity (see formal definition in Section 5), evaluated over G locations, between two given dates t_1 and t_2.

- **Multitask Lasso** (Tibshirani 1996; Jalali, Ravikumar, and Sanghavi 2013). It assumes $\mathbf{y}_{G,t} = X_t\Theta^* + \epsilon$, where $\epsilon \in \mathbb{R}^G$ is a Gaussian noise vector and $\Theta^* \in \mathbb{R}^{p\times G}$ is the coefficient matrix for all locations. With n samples, Θ^* is estimated by $\hat{\Theta}_n = argmin_{\Theta\in\mathbb{R}^{p\times G}}\frac{1}{2n}\|Y - X\Theta\|_2^2 + \lambda_n\|\Theta\|_{2,1}$ with $X \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^{n\times G}$. λ_n is a penalty parameter and the corresponding penalty term is computed as $||\Theta||_{2,1} = \sum_i(\sum_j \Theta_{ij}^2)^{1/2}$.
- **Gradient boosted trees (XGBoost)** (Friedman 2001; Chen and Guestrin 2016). A functional gradient boosting algorithm using regression tree as its weak learner. The algorithm starts with one weak learner and iteratively adds new weak learners to approximate functional gradients. The final ensemble model is constructed by a weighted summation of all weak learners.
- **State-of-the-art climate baselines.** We consider two baselines from climate science perspective, both are Least Square (LS) linear regression models (Weisberg 2005). The first model uses covariates based on climate indices, such as NAO and Niño indices, which are widely used to monitor ocean conditions. The covariate of the second model is the most recent available data point from target variable, i.e, residualized temperature of week -2 & -1, with which the model, also known as *damped persistence* (Van den Dool 2007) in climate science, is essentially a first-order autoregressive model.

DL models. We design two DL models, namely Encoder (LSTM)-Decoder (FNN) and CNN-LSTM, specifically adapting to SSF. The objective function is to minimize the mean squared error among all dates and locations.

- **Encoder (LSTM)-Decoder (FNN).** Inspired by Autoencoder widely used in sequential modeling (Sutskever, Vinyals, and Le 2014), we design the Encoder (LSTM)-Decoder (FNN) model, of which the architecture is shown in Figure 2(a). Input of the model is features extracted spatially from covariates using unsupervised methods like Principal Component Analysis (PCA). The temporal components of covariates are handled by feeding features of each historical date into an LSTM Encoder recurrently. Then, the output of each date from LSTM is sent jointly to a two-layer FNN network with ReLU activation function. The output of the FNN Decoder is the predicted residualized temperature of week 3 & 4 over all target locations.
- **CNN-LSTM.** The proposed CNN-LSTM model directly learns the representations from the spatiotemporal data using convolutional layers (LeCun et al. 1998). Shown in Figure 2(b), CNN extracts features for each climate variable at all historical dates separately. Then, the extracted features from the same date are collected and fed into an LSTM model recurrently. The temperature prediction for all target locations is done by an FNN layer taking the output of the LSTM's last layer from the latest input.

5 Data and Experimental Setup

Data description. Climate agencies across the world maintain multiple datasets with different formats and resolutions. We construct the SSF dataset by collecting climate variables (Table 1) from a diverse collection of data sources and converting them into a consistent format. In particular, temporal variables, e.g., Niño indices, are interpolated to a daily resolution, and spatiotemporal variables are interpolated to a spatial resolution of $0.5°$ by $0.5°$.

Preprocessing. Spatiotemporal climate variables are normalized by z-scoring at each location and each date using the mean and standard deviation of the corresponding day of the year over 1986-2016. Temporal covariates, e.g., Niño indices, are directly used without normalization. CNN and CNN-LSTM take the temporal and normalized spatiotemporal variables as input. Models other than CNN based mod-

Model	Mean(se)	Median (se)	0.25 quantile (se)	0.75 quantile (se)
Temporally Global Dataset				
XGBoost - one day	**0.3044(0.03)**	**0.3447(0.05)**	**0.0252(0.05)**	**0.5905(0.04)**
Lasso - one day	0.2499(0.04)	0.2554(0.06)	-0.0224(0.05)	0.5604(0.06)
Encoder (LSTM)-Decoder (FNN)	**0.2616 (0.04)**	**0.2995 (0.07)**	**-0.0719 (0.06)**	**0.6310 (0.05)**
FNN	0.0792(0.01)	0.0920(0.02)	0.0085(0.02)	0.1655(0.02)
CNN	0.1688(0.04)	0.2324(0.06)	-0.0662(0.06)	0.4768(0.04)
CNN-LSTM	0.1743(0.04)	0.2867(0.06)	-0.1225(0.07)	0.5148(0.04)
LS with NAO & Niño	**0.2415(0.03)**	**0.3169(0.04)**	**0.0454(0.05)**	**0.4624(0.03)**
Damped persistence	0.2009(0.04)	0.2310(0.06)	-0.0884(0.06)	0.5335(0.05))
MultiLLR	0.0684 (0.03)	0.1046 (0.05)	-0.1764 (0.06)	0.3156 (0.04)
AutoKNN	0.1457 (0.03)	0.1744 (0.05)	-0.1018 (0.06)	0.4000 (0.04)
Temporally Local Dataset				
XGBoost - one day	0.1965(0.04)	0.2345(0.05)	-0.0636(0.06)	0.5178(0.05)
Lasso - one day	0.1631(0.04)	0.2087(0.06)	-0.1178(0.05)	0.5059(0.05)
Encoder (LSTM)-Decoder (FNN)	0.1277 (0.04)	0.1272 (0.06)	-0.1558 (0.06)	0.4971 (0.06)

Table 2: Comparison of spatial cosine similarity of tmp2m forecasting for test sets over 2017-2018. XGBoost and Encoder (LSTM)-Decoder (FNN) have the best performance. Models achieve better performance using temporally global set compared to temporally local set.

els, e.g., XGBoost and Multitask Lasso, can not directly use spatiotemporal covariates due to the extremely high dimensionality of such covariates. In those cases, we extract the top 10 principal components (PCs) of each spatiotemporal covariate, based on PC loadings from 1986 to 2016 (for details, refer to Appendix B), and normalize PCs by z-scoring at each day of the year. The target variable is the residualized 2m temperature over the contiguous U.S. via the same normalization as spatiotemporal climate variables.

Feature set construction. We combine the PCs of spatiotemporal covariates with temporal covariates into a sequential feature set, which consists not only covariates of the target date, but also covariates of the 7^{th}, 14^{th}, and 28^{th} day prior to the target date, as well as the day of the year of the target date in the past 2 years and both the historical past and future dates around the day of the year of the target date in the past 2 years (see Appendix B for a detailed example).

Evaluation pipeline. Predictive models are created independently for each month in 2017 and 2018. To mimic a live system, we generate 105 test dates during 2017-2018, one for each week, and group them into 24 test sets by their month of the year. Given a test set, our evaluation pipeline consists of two parts: (1) "5-fold" training-validation pairs for hyper-parameter tuning, based on a "sliding-window" strategy designed for time-series data. Each validation set consists of the data from the same month of the year as the test set, and we create 5 such sets from dates in the past 5 years (2012 - 2016). Their corresponding training sets contain 10 years of data before each validation set; (2) the training set, including 30-year data in the past. To assure no overlap between the training and test set, we enforce the training set to end 28 days before the first date in the test set. More detailed explanations are included in Appendix B.

Evaluation metrics. Forecasts are evaluated by cosine similarity, the only metrics used in the Sub-Seasonal Climate Forecast Rodeo (Raff et al. 2017). The cosine similarity between $\hat{\mathbf{y}}$, a vector of predicted values, and $\mathbf{y}^*$, the corresponding ground truth, is computed as $\cos(\hat{\mathbf{y}}, \mathbf{y}^*) = \frac{\langle \hat{\mathbf{y}}, \mathbf{y}^* \rangle}{\|\hat{\mathbf{y}}\|_2 \|\mathbf{y}^*\|_2}$, where $\langle \hat{\mathbf{y}}, \mathbf{y}^* \rangle$ denotes the inner product between the two vectors. Then, the spatial cosine similarity is defined as $\cos(\hat{\mathbf{y}}_{G,t}, \mathbf{y}^*_{G,t})$, measuring the prediction skill at a date t. The temporal cosine similarity, assessing the prediction skill at a location g, is defined as $\cos(\hat{\mathbf{y}}_{g,T}, \mathbf{y}^*_{g,T})$.

6 Experimental Results

We compare the predictive skills of 10 ML models on SSF. In addition, we discuss a few aspects that impact the ML models the most, as well as the evolution of our DL models.

6.1 Results of All Methods

Temporal results. Table 2 lists the mean, the median, the 0.25 quantile, the 0.75 quantile, and their corresponding standard errors of spatial cosine similarity of all methods. Results based on relative R^2 are included in Appendix C. XGBoost, Encoder (LSTM)+Decoder (FNN) and Lasso accomplish higher predictive skills than other presented methods and can outperform climatology and two climate baseline models, i.e., LS with NAO & Niño, and damped persistence. Overall, XGBoost achieves the highest predictive skill in terms of both the mean and the median, demonstrating its predictive power. Surprisingly, linear regression with a proper feature set has good predictive performance. Even though DL models are not the obvious winner, with careful architectural selections, they still show encouraging results.

Spatial results. Figure 3 shows the temporal cosine similarity of all methods evaluated on test sets described in Section 5. Among all methods, XGBoost and the Encoder (LSTM)-Decoder (FNN) achieve the overall best performance, regarding the number of locations with positive temporal cosine similarity. Qualitatively, coastal and south regions in general are easier to predict compared to inland regions (e.g., Midwest), which might be explained by the influence of the slow-moving component, i.e., Pacific and Atlantic Ocean. Such component exhibits inertia or memory, in which anomalous condition can take relatively long period of time to decay. However, each model has its own favorable

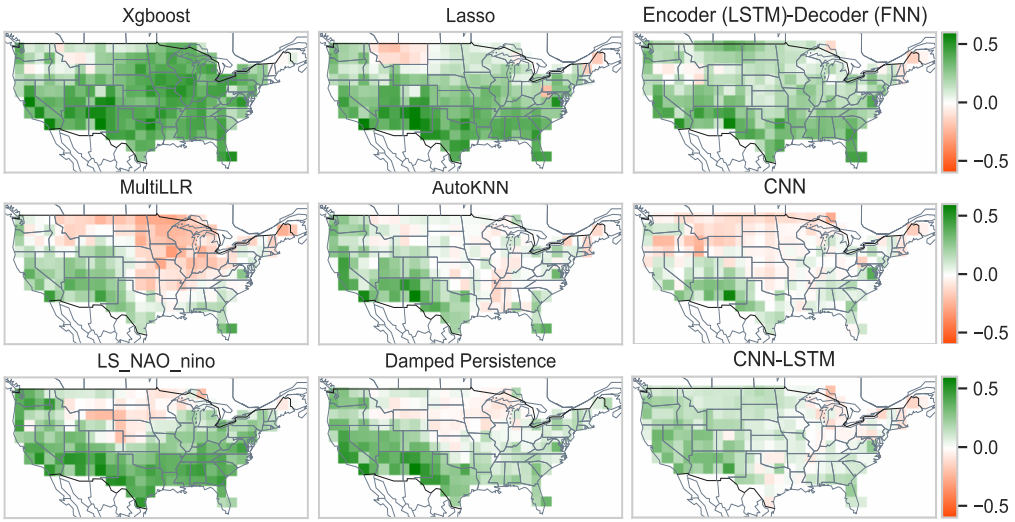


Figure 3: Temporal cosine similarity over the contiguous U.S. of ML models for temperature prediction over 2017-2018. Large positive values (green) closer to 1 indicate better predictive skills. Overall, XGBoost and Encoder (LSTM)-Decoder (FNN) perform the best. Qualitatively, coastal and south regions are easier to predict than inland regions (e.g., Midwest).

and disadvantageous regions. For example, XGBoost and Lasso do poorly in Montana, Wyoming, and Idaho, while Encoder (LSTM)-Decoder (FNN) performs much better on those regions. The observations naturally imply that the ensemble of multiple models is a promising future direction.

Comparison with the state-of-the-art methods. MultiLLR and AutoKNN are two state-of-the-art methods designed for SSF on western U.S. (Hwang et al. 2019). Both methods have shown good forecasting performance on the original target region. However, over the inland region (Midwest), Northeast, and South region, the methods do not perform so well (Figure 3). To be fair, even though a similar set of climate variables have been used in our work compared to the original paper (Hwang et al. 2019), how we prepossess the data and construct the feature set are slightly different. Such differences may lead to relatively poor performance for these two methods, especially for MultiLLR. A detailed comparison over western U.S. and on SubseasonalRodeo dataset (Hwang et al. 2019) can be found in Appendix C.

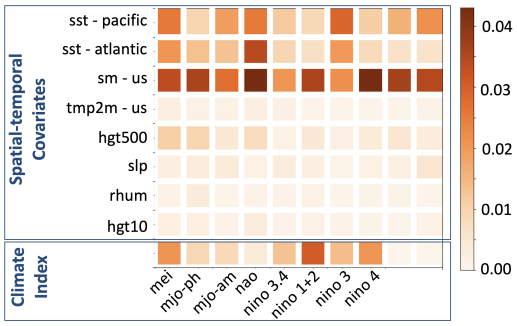
6.2 Analysis and Exploration

We analyze and explore several important aspects that could influence the performance of ML models.

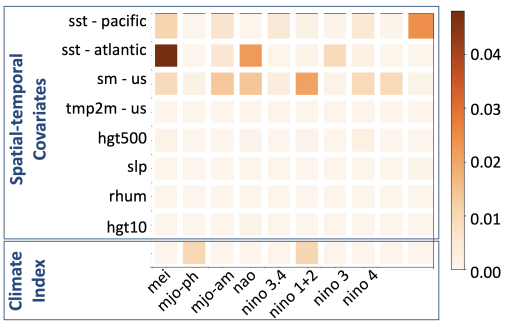
Temporally "local" vs. "global" dataset. Our current training set consists of all calendar months over the past 30 years, which we refer to as the temporally "global" dataset. Another way to construct the training set is to only consider calendar months within a temporal neighborhood of the test date. For instance, to make forecasts of June 2017, the training set can contains dates in June (from earlier years), and months that are close to June, e.g., April, May, July, and August, over the past 30 years only. Such a construction accounts for the seasonal dependence of predictive relations, for example summer predictions are not trained with winter data. We name such dataset as a temporally "local" dataset. A comparison between the "global" and "local" datasets has been listed in Table 2 where a significant drop in cosine similarity can be noticed when using "local" dataset for all of our best predictive models, including XGBoost, Lasso, and Encoder (LSTM)-Decoder (FNN). We suspect such performance drop from "global" to "local" dataset may come from the reduction in the number of effective samples.

Feature importance. We study which covariates are important, considered by ML models, based on their SHAP (SHapley Additive exPlanations) values (Lundberg and Lee 2017). SHAP values illustrate how much each feature contributes to the forecasts. Therefore, features with large absolute SHAP values are important. Figure 4 shows the mean of absolute SHAP values for each covariate over 24 models (one per month in 2017-2018), computed from (a) XGBoost and (b) Lasso. Among all covariates, soil moisture (3^{rd} row from the top) is the variable that has been constantly considered as important covariates by both models. Another set of important covariates is the family of Niño indices. An LS model using those indices alone as predictors performs fairly well (Table 2). Besides, sst of both Pacific and Atlantic also stand out. Such observations indicate that ML models pick up ocean-based covariates, some land-based covariates, and almost entirely ignore the atmosphere-related covariates, which are well aligned with domain knowledge (Uccellini and Jacobs 2018; Delsole and Tippett 2017).

The influence of feature sequence length. To adapt the usage of LSTM, we construct a sequential feature set, which consists not only the target date, but also 17 other dates preceding the target date. However, other ML models, e.g., XGBoost and Lasso, which are not designed to handle sequential data, experience a drastic performance drop when we include more information from the past. More precisely, by including covariates from the full historical sequence, the performance of XGBoost drops approximately 50% com-



(a) XGBoost



(b) Lasso

Figure 4: SHAP values computed from (a) XGBoost and (b) Lasso. Darker color means a covariate is of the higher importance. The first 8 rows contains the top 10 principal components extracted from 8 spatiotemporal covariates respectively, and the last row includes all temporal indices. Land covariate, e.g., soil moisture and ocean covariates, e.g., sst and some climate indices, are considered more important.

pared to when using covariates from the most recent date only. A possible explanation is that, as we increase the feature sequence length, such model weights covariates from different dates exactly the same without considering temporal relationship, thus irrelevant historical information might mislead the model. In Appendix C, we compares results obtained from various sequence lengths.

6.3 What Happened with DL Models?

While applying black-box DL models naively does not work well for SSF, the improvement (Table 2), as we evolve from FNN to CNN-LSTM, and finally to Encoder (LSTM)-Decoder (FNN), demonstrates how the network architecture plays an important role. Below we focus on discussing feature representation and the architecture design for sequence modeling. More discussions are included in Appendix C.

Feature representation: CNN vs. PCA. Since SSF can be considered as a spatiotemporal prediction problem, to handle the spatial aspect, CNN (LeCun et al. 1998) can be applied as a "supervised" way for learning feature representation by viewing each climate covariate as a map. CNN, while doing convolution using a small kernel, mainly focus on spatially localized regions. However, the global dependency among climate variables restricts the effectiveness of CNN kernels on feature extraction, which explains the limited predictive skill of CNN shown in Table 2 and Figure 3. Meanwhile, PCA, termed Empirical Orthogonal Functions (EOF) (Von Storch and Zwiers 2001) in climate science, is a commonly used "unsupervised" feature representation method, which focuses on low-rank modeling of spatial covariance structure revealing spatial connection. By using PCs, we are including spatial and temporal information about the dominant components of variability in each spatiotemporal covariate. Our results (Table 2) illustrate that PCA-based models have higher predictive skills than CNN-based models, verifying that PCA is an adequate technique for feature extraction in SSF.

Sequential modeling: Encoder-Decoder. With features extracted by PCA, we formulate SSF as a sequential modeling problem (Sutskever, Vinyals, and Le 2014), where the input is the covariates sequence described in Section 5, and the output is the target variable. Due to the immense success in sequential modeling (Srivastava, Mansimov, and Salakhudinov 2015), the standard Encoder-Decoder, where both Encoder and Decoder are LSTM (Hochreiter and Schmidhuber 1997), is the first architecture to investigate. Unfortunately, the model does not perform well and suffers from over-fitting, possibly caused by overly complex architecture. To reduce the model complexity, we replace the LSTM Decoder with an FNN Decoder which takes only the last step of the output sequence from the Encoder. Such change leads to an immediate boost of predictive performance. However, the input of the FNN Decoder mainly contains information encoded from the latest day in the input sequence and can only embed limited amount of historical information owing to the recurrent architecture of LSTM. To further improve the performance, we adjust the connection between Encoder and Decoder, such that FNN Decoder takes every step of the output sequence from LSTM Encoder, which makes a better use of historical information. Eventually, such architecture achieves the best performance among all investigated Encoder-Decoder variants (see details in Appendix C).

7 Conclusion

In this paper, we investigate the potential to advance subseasonal climate forecasting, a challenging and understudied problem, using ML techniques. SSF is typically a high-dimensional problem on strongly spatiotemporal correlated climate data with limited number of samples. We conduct a comprehensive study of 10 ML models, including DL models, on the SSF dataset, which is constructed for SSF over the contiguous U.S. Empirical results show the gradient boosting model XGBoost, the DL model Encoder (LSTM)-Decoder (FNN), and the linear model Lasso manage to outperform forecasts based on climatology, damped persistence and climate indices. Besides, our analysis and exploration provide insight on several essential aspects to improve the SSF performance, and show that ML models are capable of picking the climate variables from important sources of predictability on sub-seasonal time scale. With this paper, we release the SSF dataset and code base publicly, which will hopefully reduce the barrier to work on SSF for the broader ML community.

Acknowledgements

The research was supported by NSF grants OAC-1934634, IIS-1908104, IIS-1563950, IIS-1447566, IIS-1447574, IIS-1422557, CCF-1451986. The authors would like to acknowledge the computing support from the Minnesota Supercomputing Institute (MSI) at the University of Minnesota.

References

Badr, H. S.; Zaitchik, B. F.; and Guikema, S. D. 2014. Application of statistical models to the prediction of seasonal rainfall anomalies over the Sahel. *Journal of Applied meteorology and climatology* 53(3): 614–636.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Conference Track Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. URL http://arxiv.org/abs/1409.0473.

Barnston, A. G.; Tippett, M. K.; L'Heureux, M. L.; Li, S.; and DeWitt, D. G. 2012. Skill of real-time seasonal ENSO model predictions during 2002–11: Is our capability increasing? *Bulletin of the American Meteorological Society* 93(5): 631–651.

Braman, L. M.; van Aalst, M. K.; Mason, S. J.; Suarez, P.; Ait-Chellouche, Y.; and Tall, A. 2013. Climate forecasts in disaster management: Red Cross flood operations in West Africa, 2008. *Disasters* 37(1): 144–164.

Chen, T.; and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 785–794.

Cofino, A. S.; Cano, R.; Sordo, C.; and Gutiérrez, J. M. 2002. Bayesian networks for probabilistic weather prediction. In *In Proceedings of the 15th Eureopean Conference on Artificial Intelligence (ECAI)*, 695–699.

Cohen, J.; Coumou, D.; Hwang, J.; Mackey, L.; Orenstein, P.; Totz, S.; and Tziperman, E. 2019. S2S reboot: An argument for greater inclusion of machine learning in subseasonal to seasonal forecasts. *Wiley Interdisciplinary Reviews: Climate Change* 10(2): e00567.

de Perez, E. C.; and Mason, S. J. 2014. Climate information for humanitarian agencies: Some basic principles. *Earth Perspectives* 1(1): 11.

DelSole, T.; and Banerjee, A. 2017. Statistical seasonal prediction based on regularized regression. *Journal of Climate* 30(4): 1345–1361.

Delsole, T.; and Tippett, M. 2017. Predictability in a changing climate. *Climate Dynamics* doi:10.1007/s00382-017-3939-8.

Fan, Y.; and van den Dool, H. 2004. Climate Prediction Center global monthly soil moisture data set at 0.5 resolution for 1948 to present. *Journal of Geophysical Research: Atmospheres* 109(D10).

Fan, Y.; and Van den Dool, H. 2008. A global monthly land surface air temperature analysis for 1948–present. *Journal of Geophysical Research: Atmospheres* 113(D1).

Frederik Nebeker. 1995. *Calculating the weather: Meteorology in the 20th century.* Elsevier.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* 1189–1232.

Funahashi, K.-i.; and Nakamura, Y. 1993. Approximation of dynamical systems by continuous time recurrent neural networks. *Neural networks* 6(6): 801–806.

Gers, F. A.; Schmidhuber, J.; and Cummins, F. 2000. Learning to Forget: Continual Prediction with LSTM. *Neural Computation* 12(10): 2451–2471.

Goncalves, A. R.; Banerjee, A.; and Von Zuben, F. J. 2017. Spatial projection of multiple climate variables using hierarchical multitask learning. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Grover, A.; Kapoor, A.; and Horvitz, E. 2015. A deep hybrid model for weather forecasting. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 379–386. ACM.

Ham, Y.-G.; Kim, J.-H.; and Luo, J.-J. 2019. Deep learning for multi-year ENSO forecasts. *Nature* 573(7775): 568–572.

He, S.; Li, X.; Sivakumar, V.; and Banerjee, A. 2019. Interpretable Predictive Modeling for Climate Variables with Weighted Lasso. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 1385–1392.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Comput.* 9(8): 1735–1780. ISSN 0899-7667. doi:10.1162/neco.1997.9.8.1735. URL http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Hwang, J.; Orenstein, P.; Cohen, J.; Pfeiffer, K.; and Mackey, L. 2019. Improving subseasonal forecasting in the western US with machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2325–2335. ACM.

Jalali, A.; Ravikumar, P.; and Sanghavi, S. 2013. A dirty model for multiple sparse regression. *IEEE Transactions on Information Theory* 59(12): 7947–7968.

Kalnay, E.; Kanamitsu, M.; Kistler, R.; Collins, W.; Deaven, D.; Gandin, L.; Iredell, M.; Saha, S.; White, G.; Woollen, J.; Zhu, Y.; Chelliah, M.; Ebisuzaki, W.; Higgins, W.; Janowiak, J.; Mo, K. C.; Ropelewski, C.; Wang, J.; Leetmaa, A.; Reynolds, R.; Jenne, R.; and Joseph, D. 1996. The NCEP/NCAR 40-Year Reanalysis Project. *Bulletin of the American Meteorological Society* 77(3): 437–472.

Klemm, T.; and McPherson, R. A. 2017. The development of seasonal climate forecasting for agricultural producers. *Agricultural and forest meteorology* 232: 384–399.

Kunsch, H. R. 1989. The jackknife and the bootstrap for general stationary observations. *The annals of Statistics* 1217–1241.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11): 2278–2324.

Lorenc, A. C. 1986. Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society* 112(474): 1177–1194.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, 4765–4774.

National Academies of Sciences. 2016. *Next generation earth system prediction: strategies for subseasonal to seasonal forecasts*. National Academies Press.

National Research Council. 2010. *Assessment of intraseasonal to interannual climate prediction and predictability*. National Academies Press.

Pomeroy, J.; Gray, D.; Hedstrom, N.; and Janowicz, J. 2002. Prediction of seasonal snow accumulation in cold climate forecasts. *Hydrological Processes* 16(18): 3543–3558.

Radhika, Y.; and Shashi, M. 2009. Atmospheric temperature prediction using support vector machines. *International journal of computer theory and engineering* 1(1): 55.

Raff, D.; Nowak, K.; Cifelli, R.; Brekke, L. D.; and Webb, R. S. 2017. Sub-Seasonal Climate Forecast Rodeo. In *AGU Fall Meeting*.

Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; et al. 2019. Deep learning and process understanding for data-driven Earth system science. *Nature* 566(7743): 195–204.

Reshef, D. N.; Reshef, Y. A.; Finucane, H. K.; Grossman, S. R.; McVean, G.; Turnbaugh, P. J.; Lander, E. S.; Mitzenmacher, M.; and Sabeti, P. C. 2011. Detecting novel associations in large data sets. *science* 334(6062): 1518–1524.

Reynolds, R. W.; Smith, T. M.; Liu, C.; Chelton, D. B.; Casey, K. S.; and Schlax, M. G. 2007. Daily high-resolution-blended analyses for sea surface temperature. *Journal of Climate* 20(22): 5473–5496.

Schneider, T.; Lan, S.; Stuart, A.; and Teixeira, J. 2017. Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. *Geophysical Research Letters* 44(24): 12–396.

Shi, X.; Gao, Z.; Lausen, L.; Wang, H.; Yeung, D.-Y.; Wong, W.-k.; and Woo, W.-c. 2017. Deep learning for precipitation nowcasting: A benchmark and a new model. In *Advances in neural information processing systems*, 5617–5627.

Simmons, A. J.; and Hollingsworth, A. 2002. Some aspects of the improvement in skill of numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society* 128(580): 647–677.

Srivastava, N.; Mansimov, E.; and Salakhudinov, R. 2015. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, 843–852.

Steinhaeuser, K.; Chawla, N. V.; and Ganguly, A. R. 2011. Complex networks as a unified framework for descriptive analysis and predictive modeling in climate science. *Statistical Analysis and Data Mining,* 4(5): 497–511.

Sundermeyer, M.; Schlüter, R.; and Ney, H. 2012. LSTM neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society,* 267–288.

Uccellini, L. W.; and Jacobs, N. G. 2018. *Subseasonal and Seasonal Forecasting Innovation: Plans for the Twenty-First Century*. National Weather Service (U.S.).

Van den Dool, H. 2007. *Empirical methods in short-term climate prediction*. Oxford University Press.

Van den Dool, H.; Saha, S.; and Johansson, A. 2000. Empirical orthogonal teleconnections. *Journal of Climate* 13(8): 1421–1435.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Venugopalan, S.; Rohrbach, M.; Donahue, J.; Mooney, R.; Darrell, T.; and Saenko, K. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, 4534–4542.

Vitart, F.; Robertson, A. W.; and Anderson, D. L. 2012. Subseasonal to Seasonal Prediction Project: Bridging the gap between weather and climate. *Bulletin of the World Meteorological Organization* 61(23).

Von Storch, H.; and Zwiers, F. W. 2001. *Statistical analysis in climate research*. Cambridge university press.

Weisberg, S. 2005. *Applied linear regression*, volume 528. John Wiley & Sons.

Wheeler, M. C.; and Hendon, H. H. 2004. An All-Season Real-Time Multivariate MJO Index: Development of an Index for Monitoring and Prediction. *Monthly Weather Review* 132(8): 1917–1932.

Xingjian, S.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; and Woo, W.-c. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, 802–810.

Zimmerman, B. G.; Vimont, D. J.; and Block, P. J. 2016. Utilizing the state of ENSO as a means for season-ahead predictor selection. *Water resources research* 52(5): 3761–3774.