Learning and Dynamical Models for Sub-seasonal Climate Forecasting: Comparison and Collaboration

Sijie He[1, 3], Xinyan Li[1], Laurie Trenary[2], Benjamin A Cash[2], Timothy DelSole[2], Arindam Banerjee[3]

[1]Department of Computer Science & Engineering, University of Minnesota, Twin Cities
[2] Department of Atmospheric, Oceanic, and Earth Science, George Mason University
[3] Department of Computer Science, University of Illinois Urbana-Champaign
{hexxx893, lixx1166}@umn.edu, {ltrenary, bcash, tdelsole}@gmu.edu, arindamb@illinois.edu

Abstract

Sub-seasonal forecasting (SSF) is the prediction of key climate variables such as temperature and precipitation on the 2-week to 2-month time horizon. Skillful SSF would have substantial societal value in areas such as agricultural productivity, hydrology and water resource management, and emergency planning for extreme events such as droughts and wildfires. Despite its societal importance, SSF has stayed a challenging problem compared to both short-term weather forecasting and long-term seasonal forecasting. Recent studies have shown the potential of machine learning (ML) models to advance SSF. In this paper, for the first time, we perform a fine-grained comparison of a suite of modern ML models with start-of-the-art physics-based dynamical models from the Subseasonal Experiment (SubX) project for SSF in the western contiguous United States. Additionally, we explore mechanisms to enhance the ML models by using forecasts from dynamical models. Empirical results illustrate that, on average, ML models outperform dynamical models while the ML models tend to generate forecasts with conservative magnitude compared to the SubX models. Further, we illustrate that ML models make forecasting errors under extreme weather conditions, e.g., cold waves due to the polar vortex, highlighting the need for separate models for extreme events. Finally, we show that suitably incorporating dynamical model forecasts as inputs to ML models can substantially improve the forecasting performance of the ML models. The SSF dataset constructed for the work and code for the ML models are released along with the paper for the benefit of the artificial intelligence community.

1 Introduction

Over the past decade, good quality short-term (few days) weather forecasts as well long-term (beyond few months) seasonal forecasts have both become routinely available. These forecasts are largely based on dynamical models that solve partial differential equations (PDEs) derived from the laws of physics. In contrast, skillful sub-seasonal forecasts (SSF), i.e., the prediction of key climate variables such as temperature and precipitation on 2-week to 2-month time scales, are arguably not yet available. Skillful SSF has immense societal value as discussed in two recent reports from the National Academy of Sciences (NAS) (National Research Council 2010; National Academies of Sciences 2016). In particular, high-quality SSF in the western contiguous United States would allow for better water resource management and emergency planning for extreme events such as droughts and wildfires (White et al. 2017). Currently, sub-seasonal forecasts based on dynamical models are available weekly through the Subseasonal Experiment (SubX) project (Pegion et al. 2019), but the full utility of these for operational forecasting still remains to be determined.

SSF is challenging for a variety of reasons. First, high-quality SSF has proven difficult to accomplish compared to both short-term weather forecasting and long-term seasonal forecasting (Vitart, Robertson, and Anderson 2012). Due to the chaotic nature of atmosphere, weather events can not be accurately predicted beyond two weeks using dynamical models (Lorenz 1963). From a physical point of view, the predictability on sub-seasonal time scales depends on correctly modeling the atmosphere, ocean, and land, including their interactions and couplings as well as the memory effects of land and ocean. In addition to these physical complexities, SSF poses unconventional time series prediction problems. Given a training set $\{x_{1:t}, y_{1:t}\}$, where y denotes the target response variable, e.g., land temperature, and x denotes suitable covariates, temporal models typically focus on predicting y_{t+1} or maybe $y_{t+1:t+\tau_s}$ for a small τ_s (a few days or less). Instead, SSF is about predicting $y_{t+T:t+T+\tau_l}$ for large $T \gg \tau_s$, e.g., weather prediction one month ahead ($T = 31$ days). The long temporal range relative to the weather predictability time, along with the nonlinear dynamics and complex interactions, makes SSF challenging.

For climate forecasting, one standard baseline for comparing forecasts is the so-called climatology (Trewin et al. 2007), i.e. the 30-year average temperature/precipitation for each calendar day at each geographic location. Despite its simplicity, climatology provides a competitive benchmark for SSF. For instance, in the last Forecast Rodeo (NIDIS 2019), a SSF competition sponsored by the U.S. Bureau of Reclamation and the NOAA/National Integrated Drought Information System (USBR and NOAA 2019), about half of the submitted forecasts could not beat climatology. Thus, for any more advanced SSF models, the first order of business is to do better than climatology. Recently, progress has been made in developing ML models (Hwang et al. 2019; He et al. 2021; Weyn et al. 2021; Srinivasan et al. 2021) which have shown great promise for outperforming climatology.

In this paper, we consider two new directions for SSF: first, comparing and contrasting ML models for SSF with an arguably stronger baseline provided by physics-based dynamical models; and second, exploring enhancing the ML models by using forecasts from such dynamical models. For the comparison, earlier literature has done such comparisons with certain statistical approaches and has illustrated dynamical models to have better forecasting ability (Barnston et al. 2012). Instead, we do the comparison with a suite of modern ML methods, including non-parametric AutoKNN (Hwang et al. 2019), multitask Lasso (Tibshirani 1996; Jalali, Ravikumar, and Sanghavi 2013), gradient boosted trees (Friedman 2001; Chen and Guestrin 2016), and deep encoder-decoder networks (He et al. 2021), and illustrate that on average ML models outperform dynamical models on SSF. With considerably more details, our empirical analysis demonstrates key properties of ML-based vs. dynamical model-based predictions. In particular, most ML models generate conservative forecasts with small values, whereas dynamical models are more aggressive, generating forecasts with large scale. So when dynamical models are wrong, they can be wrong by a large amount; on the flip side, when dynamical models are correct, they can be more accurate than ML models. Further, we illustrate that ML models make most of their bad predictions during extreme events, e.g., unusual cold waves in North America, for which there is not enough training data. These results suggest that a separate ML model for extreme events will potentially help improve aggregate performance. The second direction is using physics-based dynamical model forecasts as covariates in the ML models. We show that using dynamical model forecasts as inputs improves the ML model forecasts, and the improvements are statistically significant. In addition to the extensive new results on SSF, we release all the data and code to replicate and hopefully extend our work[1]. We want to enable a new application area for artificial intelligence research, focusing on a challenging and societally important scientific problem in the context of climate forecasting.

2 Related Work

Dynamical models and S2S forecasting. Nowadays, weather predictions rely heavily on ensemble forecasts from physics-based dynamical models (Barnston et al. 2012). On sub-seasonal to seasonal (S2S) time scales, forecasts have shown limited predictive skill compared to the climatology (Vitart 2004, 2014; Weigel et al. 2008). However, successful S2S predictions can be performed for certain regions and seasons (Li and Robertson 2015; DelSole et al. 2017a), as well as certain climate states (Mariotti et al. 2020). To understand the conditions that lead to enhance predictability and to improve S2S forecasts, projects such as S2S (Vitart et al. 2017) and SubX (Pegion et al. 2019) have been established. These coordinated multi-model efforts act to fulfill the growing needs of skillful SSF in real-world applications.

ML on weather and S2S forecasting. Recently, increasing efforts have been made to tackle complex problems in climate science using ML. Such applications aim to advance weather forecast skill using deep learning methods (Liu et al. 2016; Ham, Kim, and Luo 2019; Dueben and Bauer 2018; Scher and Messori 2019). Despite early studies that show dynamical models outperform statistical models for ENSO seasonal forecasts (Barnston et al. 2012), recent advances in machine learning, especially the development of deep learning, are making the performance of ML models more competitive with dynamical models for both weather (Grover, Kapoor, and Horvitz 2015; Shi et al. 2017; Dueben and Bauer 2018) and seasonal (Stevens et al. 2021) prediction.

In particular, ML models have started to be used to improve forecast skills for predictions of temperature, precipitation, and other climate variables on sub-seasonal time scales (Hwang et al. 2019; He et al. 2021; Weyn et al. 2021; Srinivasan et al. 2021). Some successful ML approaches for S2S forecasting include (Hwang et al. 2019) and (He et al. 2021), where both works show increased predictive skill for ML models compared to climatic baselines, e.g., climatology and damped persistence. Such advances from ML models are particularly relevant and valuable, as dynamical models have limited predictive skills at sub-seasonal time scales (Uccellini and Jacobs 2018).

3 Sub-seasonal Climate Forecasting

Problem Statement. We focus on forecasting temperature anomalies over days 15 - 28, i.e., predicting average temperatures anomalies 2 weeks ahead of time, over the western contiguous U.S, which follows the Forecast Rodeo competition (USBR and NOAA 2019). The spatial region is bounded by latitudes 25N-50N and longitudes 93W-125W at 1° by 1° spatial resolution with 508 grid points. The temporal range of interest and temporal resolutions are determined by each SubX model and its initialization frequency.

Ground Truth Dataset. The ground truth dataset is constructed from NOAA's Climate Prediction Center (CPC) Global Gridded Temperature dataset, which is commonly applied for forecast verification by NOAA/CPC (Fan and Van den Dool 2008). The CPC dataset provides daily max and min 2m temperatures (tmp2m - air temperature at 2 meters above the surface) at 0.5 ° by 0.5 ° spatial resolution from Jan 1, 1979 to the present. To obtain the ground truth temperature anomalies for weeks 3 &4, we preprocess the data as follows: (1) daily 2m temperature at each grid point is taken as the average of daily max and min tmp2m, (2) all missing values are imputed by averaging the daily tmp2m of its spatial/temporal neighbors, (3) the tmp2m at $0.5\,^{\circ} \times 0.5^{\circ}$ resolution are linearly interpolated to a $1\,^{\circ} \times 1^{\circ}$ grid, (4) the daily tmp2m anomalies are computed by subtracting the climatology from the observed daily tmp2m, and (5) the forecasting target at each date and grid point is the average of tmp2m anomalies at day 15 to 28. The climatology used in step (4) is the smoothed long-term average of tmp2m over 1990 - 2016 for each month-day combination and grid point. Specifically, for a given grid point, we compute the long-term average over 1990 - 2016, one for each month-day combination. Then the 365 values are smoothed using

[1] The SSF dataset is publicly available at https://sites.google.com/view/ssf-dataset. The source code can be found at https://github.com/Sijie-umn/SSF-MIP.

moving average with a window size of 31 days.

Evaluation Metrics. Let $\mathbf{y}^* \in \mathbb{R}^n$ denotes the ground truth observation and $\hat{\mathbf{y}} \in \mathbb{R}^n$ be the corresponding predicted value, we consider the following two evaluation metrics.

Anomaly Correlation Coefficient (ACC) is defined as $\text{ACC} = \frac{\text{cov}(\mathbf{y}^*, \hat{\mathbf{y}})}{\sigma_{\mathbf{y}^*}\sigma_{\hat{\mathbf{y}}}}$, where $\sigma_{\hat{\mathbf{y}}}$ ($\sigma_{\mathbf{y}^*}$) represents the standard deviation of $\hat{\mathbf{y}}$ ($\mathbf{y}^*$). $\text{cov}(\mathbf{y}^*, \hat{\mathbf{y}})$ is the covariance between $\mathbf{y}^*$ and $\hat{\mathbf{y}}$. ACC is independent of the mean and variance of each individual distribution of $\mathbf{y}^*$ and $\hat{\mathbf{y}}$ and is equivalent to cosine similarity, the evaluation metric used in Forecast Rodeo, assuming $\hat{\mathbf{y}}$ and $\mathbf{y}^*$ are zero-mean.

Relative R^2 is defined as $1 - \frac{\sum_{i=1}^{n}(\mathbf{y}_i^* - \hat{\mathbf{y}}_i)^2}{\sum_{i=1}^{n}(\mathbf{y}_i^* - \bar{\mathbf{y}}_{\text{train}})^2}$, where $\bar{\mathbf{y}}_{\text{train}}$ is the long-term average of tmp2m at each date and grid point in the training set. Relative R^2 is equivalent to $1-$Relative MSE and represents the relative skill against the best constant predictor, i.e., $\bar{\mathbf{y}}_{\text{train}}$. A model which achieves a positive relative R^2 is, at least, able to predict the sign of $\mathbf{y}^*$ accurately and outperforms the climatology.

Denote the ground truth temperature anomalies as $Y^* \in \mathbb{R}^{T \times G}$, where T is the number of dates and G is the number of grid points. The *spatial* predictive skill for a given date t can be evaluated on $\mathbf{y}_t^* = Y^*[t, :]$, the t-th row in Y^*, where $\mathbf{y}_t^* \in \mathbb{R}^G$ is the ground truth for all grid points at date t with the corresponding forecasts $\hat{\mathbf{y}}_t$. The *temporal* predictive skill for a grid point g can be evaluated on $\mathbf{y}_g^* = Y^*[:, g]$, the g-th column in Y^*, similar to time series prediction evaluation.

4 Subseasonal Experiment (SubX) Project

The Subseasonal Experiment (SubX) project provides subseasonal forecasts from multiple global forecast models. Data are publicly available through the International Research Institute for Climate and Society (IRI) Data Library at Columbia University. A detailed description of the SubX project and the contributing models can be found in (Pegion et al. 2019). The SubX project has two predictive periods: hindcast and forecast. A hindcast period represents the time when a dynamic model re-forecasts historical events, which can help climate scientists develop and improve forecasting models. In contrast, a forecast period has real-time predictions generated from dynamic models. Specifically, hindcasts occur from Jan 1999 to Dec 2015, while the real-time forecast period starts from July 2017. We evaluate the predictive skills of the SubX models over their forecast periods.

In this paper, we focus on two SubX models, NCEP-Climate Forecast System version 2 (CFSv2) (Saha et al. 2014) and NASA-Global Modeling and Assimilation (GMAO) version 2 of the Goddard Earth Observing System (GEOS) model (Reichle and Liu 2014). NCEP-CFSv2 is the operational seasonal prediction model currently used by the U.S. Climate Prediction Center. GMAO-GEOS is developed to support NASA's earth science research. Both models are coupled atmosphere–ocean–land–sea ice models and have the highest initialization frequency in the SubX project. Further information on the two SubX models is presented in the Appendix[2]. For each SubX model, there are four ensemble members and daily forecasts for 45 days beyond each initialization date. The average of four ensemble members' outputs are taken as the forecasts. The weeks 3 & 4 outlooks are computed by averaging the forecasts 15 to 28 days beyond each initialization date and subtracting the corresponding climatology computed from the model's hindcast period.

5 Machine Learning-based SSF Modeling

Notation. Let $Y \in \mathbb{R}^{T \times G}$ denote the targeted weeks 3 & 4 temperature anomalies over T dates and G grid points. $\mathbf{y}_t$ is the t-th row in Y, denoting the temperature anomalies over all grid points G at date t. $X \in \mathbb{R}^{T \times p}$ denotes the p-dimension covariates for T dates. $X_t \in \mathbb{R}^p$ (the t-th row in X) is the covariates at date t.

Machine Learning Models. We focus on state-of-the-art machine learning models that have been shown to work effectively for SSF (He et al. 2021; Hwang et al. 2019).

AutoKNN (Hwang et al. 2019). An auto-regressive model only uses features from historical temperature anomalies, which are selected using a multitask k-nearest neighbor criterion. For a given date t, the algorithm chooses the temperature anomalies of 20 historical dates which have the highest similarity with the date t, and temperature anomalies of 29 days, 58 days, and 1 year prior to t as features. Specifically, the similarity between two dates t_1 and t_2 is defined as $\text{sim}_{(t_1,t_2)} = \frac{1}{M}\sum_{m=0}^{M-1}\cos(\mathbf{y}_{t_1-l-m}, \mathbf{y}_{t_2-l-m})$, where $\cos(\mathbf{y}_{t_1-l-m}, \mathbf{y}_{t_2-l-m})$ is the cosine similarity between the temperature anomalies at $l+m$ days before t_1 and t_2. Following the settings in (Hwang et al. 2019), we use $M = 60$ as the length of the considered historical sequences prior to each date, with the lag $l = 365$. At each grid point, we fit a weighted local linear regression model, where the weight is one over the variance of the temperature anomalies at the corresponding date.

Multitask Lasso (Tibshirani 1996; Jalali, Ravikumar, and Sanghavi 2013). A multitask regularized linear regression model. By assuming $\mathbf{y}_t = X_t\Theta^* + \epsilon$, where $\epsilon \in \mathbb{R}^G$ is a Gaussian noise and $\Theta^* \in \mathbb{R}^{p \times G}$ is the coefficient matrix for all locations, the parameter Θ^* is estimated by

$$\hat{\Theta} = \text{argmin}_{\Theta \in \mathbb{R}^{p \times G}} \frac{1}{2T}\|Y - X\Theta\|_2^2 + \lambda\|\Theta\|_{2,1} \qquad (1)$$

with $||\Theta||_{2,1} = \sum_i(\sum_j \Theta_{ij}^2)^{1/2}$ and a penalty parameter λ .

Gradient boosted trees (XGBoost) (Friedman 2001; Chen and Guestrin 2016). A functional gradient boosting algorithm, of which the weak learners are regression trees. The algorithm combines multiple weak learners into one learner in an iterative manner. At each iteration, a new weak learner is created to correct the previous prediction and optimize the loss function along with regularization. We build one XGBoost model for each location, and the hyper-parameters are selected jointly based on the performance over all locations.

Encoder-FNN (He et al. 2021). A deep learning model designed for SSF over the contiguous U.S. The model input is a historical sequence of the features shared by all locations and is fed into an LSTM encoder recurrently. The outputs of each step in the sequence are combined and jointly sent to the decoder, which is a two-layer fully-connected neural network with ReLU activation. The outputs of the decoder

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

are the predicted tmp2m anomalies over all grid points. Note that, besides standard hyper-parameters like layer size, number of layers, and dropout rate, the length of the sequence is also a hyper-parameter. The final forecast is the average of 20 independent runs.

Covariates for ML models. The feature set for the ML models contains the following climate variables, of which the detailed description and original sources are listed in the Appendix. Spatially over the contiguous U.S., we consider (1) 2m temperature (tmp2m), which is also the source data for the ground truth dataset; (2) soil moisture (sm), which influences temperature and precipitation through its impact on surface fluxes of heat and moisture (Koster et al. 2011); and (3) geopotential height (ght) at 10mb and 500mb, sea level pressure (slp) and relative humidity (rhum) from the reanalysis dataset, which capture variations in the northern hemisphere polar vortex and persistent variations in the large-scale atmospheric circulation. We also obtain sea surface temperature (sst) over the Pacific Ocean, from latitudes 20S to 65N and longitudes 120E to 90W, and the Atlantic Ocean, from latitudes 20S to 50N and longitudes 20W to 90W. Variations in sst have been linked to enhanced subseasonal predictability over the U.S. (DelSole et al. 2017b).

In addition, we include nine climate indices that describe the state of the climate system or are related to different climate phenomena, such as El Niño/Southern Oscillation (ENSO). Multivariate ENSO index (MEI.v2) and Niño indices are included to monitor El Niño and La Niña events (DelSole et al. 2017b; Stan et al. 2017). The amplitude and phase of Madden-Julian Oscillation (MJO) are considered since the MJO has dramatic impacts in the mid-latitudes and is a strong contributor to various extreme events in the U.S. (Waliser 2005). North Atlantic Oscillation (NAO) index is considered since variations in the NAO drive changes in temperature and precipitation over the U.S. and western Europe (Stan et al. 2017). Sudden Stratospheric Warming (SSW) index is included to capture the variations in the strength of the polar vortex, which are associated with extreme cold air outbreaks in mid-latitude U.S. (Butler et al. 2015).

Data Preprocessing. For all ML models except AutoKNN, we consider two types of climate variables, namely spatiotemporal and temporal climate variables. For each spatiotemporal variable, we flatten the values at all grid points for each date and compute the top 10 principal components (PCs) as features. For example, if $X^{\text{sm}} \in \mathbb{R}^{T \times G}$ denotes the soil moisture for T dates in training set (1990-2016) and all G spatial grid points over the contiguous U.S., we compute the PC loadings using X^{sm} and extract the top 10 PCs to get the feature matrix $X^{\text{sm}}_{\text{pc}} \in \mathbb{R}^{T \times 10}$. The extracted PCs are then normalized by z-scoring for each month-day combination separately. The temporal variables and the PC-based features of all spatiotemporal climate variables jointly form the feature set for each date. For XGBoost and Lasso, the covariates are the feature values two weeks lagging from the forecasting period. For example, if the forecasting period is Jan, 15 - Jan, 28 in 2019, the covariates are the features on Jan 1, 2019. For Encoder FNN, the features of a historical sequence are treated as the model input for each date. The historical sequence is constructed similarly to the features of Encoder FNN in (He et al. 2021) (Figure 7(a)). AutoKNN takes only the historical tmp2m anomalies as its covariate.

Experimental Setup. Since the relationships between the covariates and target variables vary at different times of the year, test sets are created for each month from July 2017 to Jun 2020 and separate predictive models are trained accordingly. Since an individual ML model is built for each month of the year, the best hyper-parameters of each type of ML models are selected on a monthly basis. To do so, for each month of the year, we construct five validation sets containing data from the same month between 2012 and 2016, and the corresponding training sets consist of 10 years of data prior to each validation set. The best hyper-parameters are determined by the average performance over the five validations sets. We thus have 12 sets of the best hyper-parameters corresponding to each month of the year. Once the best hyper-parameters are selected, we use 28 years of data prior to a given test set to train the corresponding ML forecasting model.

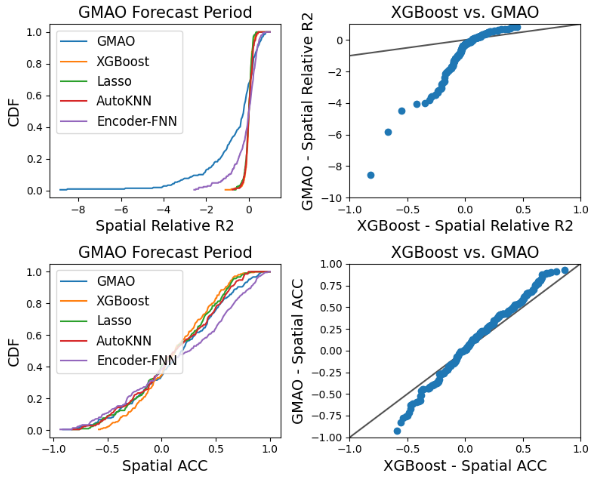


Figure 1: The empirical cumulative distribution function (cdf) of spatial relative R^2 (top left) and spatial ACC (bottom left) of all methods, and the corresponding quantile-quantile (QQ) plot (right) between XGBoost and GMAO-GEOS. XGBoost, Lasso and AutoKNN all have most spatial relative R^2 close to or above 0, while GMAO-GEOS and Encoder-FNN have relative R^2 much smaller than -1. As for spatial ACC, despite the similarity of the cdf curves, the ML models (yellow, green, and red) are in general below the blue curve when the spatial ACCs are negative, indicating that the ML models are less likely to have extremely negative predictive skills compared to the SubX model.

6 Experimental Results

In this section, we compare the predictive skill of the four ML models and the two SubX models on the forecast period from 2017 to 2020. A comprehensive analysis is conducted for the experimental results, which reveals possible directions for further improvement of the ML models for SSF. Besides, we explore the potential of advancing SSF by combining the ML models and the SubX forecasts.

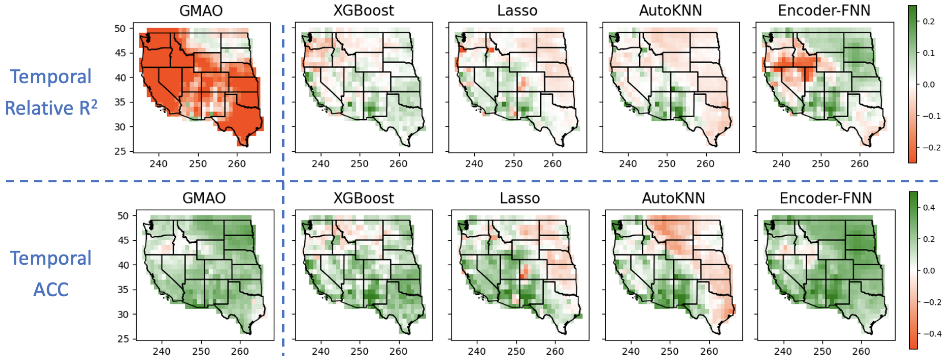


Figure 2: Temporal relative R^2 (top) and temporal ACC (bottom) of GMAO-GEOS (leftmost column) and the ML models. For both metrics, values closer to 1 (green) indicate more accurate predictions. Overall the SubX model achieves positive temporal ACC for most spatial locations while performing poorly considering temporal relative R^2. Among all the ML models, XGBoost and Encoder-FNN are the two best models, and substantially outperform the SubX models over most spatial locations.

Forecast Period Evaluation

We evaluate the results for GMAO-GEOS and NCEP-CFSv2 separately since they have different forecast periods and temporal resolutions. Overall the results are consistent for both of the SubX models, therefore we present only the results regarding GMAO-GEOS in this section and the results regarding NCEP-CFSv2 in the Appendix. We first present the empirical cumulative distribution function (cdf) of spatial relative R^2 and spatial ACC for all methods over the forecast periods of GMAO-GEOS in Figure 1. It is shown that ML models such as XGBoost, Lasso, and AutoKNN are capable of generating forecasts with positive or small negative relative R^2, while the SubX model and the Encoder-FNN model commonly stay in the negative relative R^2 zone. On the other hand, considering the positive side of the cdf plot, the SubX model and Encoder-FNN are able to achieve relative R^2 close to 1 in some cases, whereas the cdf of other ML models reach 1 when the relative R^2 are comparatively small. The quantile-quantile (QQ) plot of spatial relative R^2 in Figure 1 shows that the relative R^2 can be much smaller than -1, indicating the SubX models can make predictions with a large deviation from the ground truth. As for spatial ACC, despite the similarities in the cdf across models, a closer inspection shows the cdf of the ML models (yellow, green, and red curves) are generally below the cdf of the SubX model (blue curve) between [-1, 0]. The QQ plot of the spatial ACC between XGBoost and GMAO-GEOS supports the observation, where all the points are below the diagonal line when the spatial ACC of XGBoost is between [-1, 0]. For the positive side of the spatial ACC for XGBoost, most points are close to or slightly above the diagonal line. To summarize, at a given date, the SubX model is more likely to have spatial ACC close to the extreme values (-1 or +1), while ML models, such as XGBoost, are more conservative and are able to avoid extreme negative ACC.

The temporal relative R^2 and temporal ACC over the western U.S. are illustrated in Figure 2. Similar to spatial results, the SubX model achieves positive temporal ACC for most spatial locations while performing poorly with respect to temporal relative R^2. Among all ML models, XGBoost and Encoder-FNN are the best two considering both temporal predictive skills, and substantially outperform the SubX model for most spatial locations. Spatially, the central area, including the states of North Dakota, South Dakota, Montana, Wyoming, etc, are the areas where the temperature fluctuations are more drastic compared to the coastal states. Therefore, linear model like Lasso and non-parametric model like AutoKNN tend to perform worse in such regions, while more complicated nonlinear models like XGBoost and Encoder-FNN perform relatively better. Additionally, the SubX model has negative temporal relative R^2 and positive temporal ACC for the coastal area, which implies it may predict incorrect magnitudes despite their relatively accurate prediction of the temporal patterns.

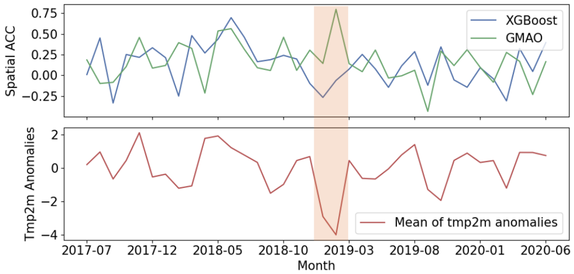


Figure 3: The monthly average spatial ACC of XGBoost and GMAO-GEOS and the mean of tmp2m anomalies over the western U.S. during the forecast period. Most of the time, XGBoost achieves competitive or even higher spatial ACC compared to the SubX model. The only exception, that the SubX model (same for NCEP-CFSv2) significantly outperform XGBoost, happens from Dec. 2018 to Feb. 2019 (highlighted in orange) when a cold wave affected the U.S. leading to extreme low average tmp2m anomalies.

Machine Learning and Extreme Weather Events

Given that SSF is a challenging problem, it is natural to investigate under which circumstance(s) the ML models fail to

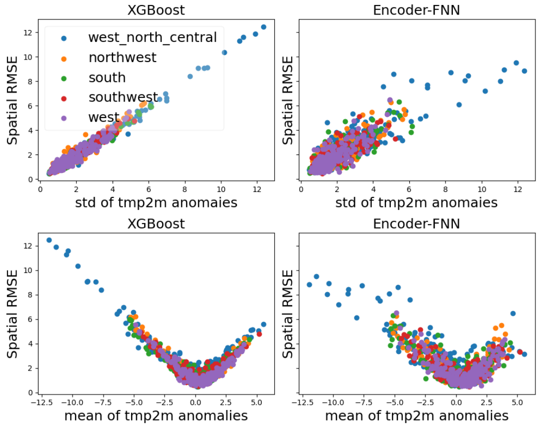
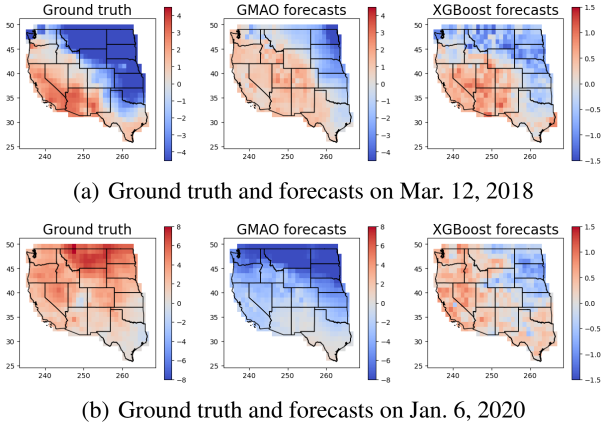


Figure 4: Spatial RMSE versus the standard deviation of tmp2m anomalies (top) and the average tmp2m anomalies (bottom) over the dates and regions. The high spatial RMSE appears for samples having large standard deviation or extreme negative average of tmp2m anomalies, which indicates that the west-north-central region is hard to predict.

provide accurate forecasts. The average spatial ACC of XGBoost and GMAO-GEOS for each month during the forecast periods are shown in Figure 3. For most months, XGBoost is either competitive or achieves higher spatial ACC compared to the SubX model. The exceptions occur in December 2018 and the first two months of 2019, when the January–February 2019 North American cold wave impacted the United States. The cold wave brought the coldest temperatures in over 20 years to most locations (Wikipedia 2019), and the temperature anomalies reached -15°C and beyond in the *central U.S.* Extreme weather events are hard to predict since there is a lack of enough training data for such events. However, the dynamical models are reasonably successful in predicting the extreme cold temperatures, since they follow the physics. For example, the cold wave followed a sudden stratospheric warming event, which increases predictability of these extreme events (Domeisen and Butler 2020).

The value of spatial ACC is not affected by the scale of the response. Therefore, we also analyze the predictive performance for the ML models regarding Root Mean Square Error (RMSE). With $\mathbf{y}_i^*$ and $\hat{\mathbf{y}}_i$ denoting the i-th element in the ground truth $\mathbf{y}^*$ and the forecasts $\hat{\mathbf{y}}$ respectively, RMSE is defined as $\sqrt{\frac{\sum_{i=1}^{n}(\mathbf{y}_i^* - \hat{\mathbf{y}}_i)^2}{n}}$. We first separate all grid points in the western U.S. into five climatically consistent regions (Karl and Koss 1984), i.e., northwest, west, west-north-central, southwest, and south. To represent the spatial variance of tmp2m anomalies at each forecasting date and each region, we approximately compute the standard deviation (std) of tmp2m anomalies at each date and each region as $\sqrt{\sum_{i=1}^{n_r} \frac{y_i^2}{n_r}}$, where n_r is the number of grid points for a given region at one date. As shown in Figure 4, the RMSE from ML models at a given date and region is strongly correlated to the std of tmp2m anomalies, which implies the dates and regions with high variance are difficult to predict.



(a) Ground truth and forecasts on Mar. 12, 2018

(b) Ground truth and forecasts on Jan. 6, 2020

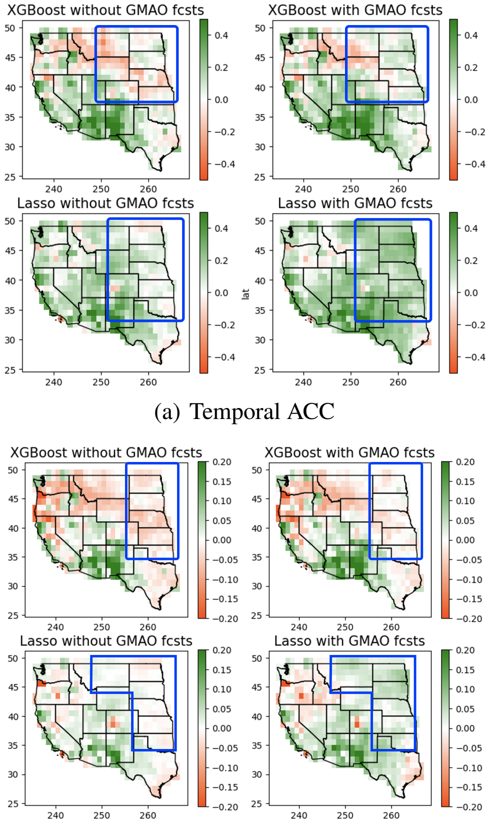
Figure 5: Comparison among the ground truth and forecasts made by GMAO-GEOS and XGBoost at two dates. (a) On March 12, 2018, both XGBoost and GMAO-GEOS successfully predict the spatial pattern of the ground truth (red in the southwest and blue in the northeast). However, the predicted values from XGBoost are much smaller than GMAO-GEOS forecasts as XGBoost is more conservative on the scale. (b) On Jan 6, 2020, the ground truth has positive tmp2m anomalies (red) for most locations, while GMAO-GEOS mistakenly makes extreme negative forecasts (dark blue).

The bottom plots in Figure 4 illustrate the average of tmp2m (with sign, unlike the std) for each date and region versus the predictive RMSE, which further demonstrates that extreme events are the samples with negative bias and large variance during the forecast period. Results for other ML models are included in the Appendix. Besides, the distribution of different regions in Figure 4 implies that the spatial variance is, in general, lower for coastal regions, e.g. west region, compared to inland regions like west-north-central region. For instance, west-north-central region (including Montana, Wyoming, North Dakota, South Dakota, and Nebraska) can experience extremely cold winter temperatures when the polar jet stream sinks down into the mid-latitudes and brings the coldest polar air.

This analysis illustrates the difficulty of modeling extreme weather events using a single ML model, not only because of the inadequate samples, but also due to the intense temperature fluctuations. Therefore, it is necessary to utilize separate modeling techniques for weather extremes or regions with drastic fluctuations in tmp2m anomalies to achieve more accurate forecasting. Ideally, if weather extremes can be detected ahead of time, we can choose not to trust the ML forecasts for a certain time period and turn to the forecasting models specifically designed for extreme conditions.

Enhancing ML Models with SubX Forecasts

To demonstrate the strengths and limitations of the SubX and the ML model forecasts, we present forecasts of two days as anecdotal evidence in Figure 5. The first example (Figure 5(a)) shows that, on Mar. 12, 2018, both GMAO-GEOS and XGBoost have successfully reproduced the spa-



(a) Temporal ACC

(b) Temporal relative R^2

Figure 6: The temporal ACC and relative R^2 of XGBoost and Lasso with and without GMAO forecasts as features. Including GMAO forecasts in the feature set evidently improves the forecasting performance, especially for the central U.S. (top right corner, marked by blue frames).

tial pattern of the ground truth. As a result, GMAO-GEOS and XGBoost obtain good spatial ACC. However, the predicted scale from GMAO-GEOS is much larger than XGBoost and is closer to the scale of the ground truth. The second example is the forecasting results on Jan 6, 2020, when the SubX forecasts fail badly. As shown in Figure 5(b), while the ground truth is that all the locations over the western U.S. have positive tmp2m anomalies with the largest values around 8°C, GMAO-GEOS predicts all negative tmp2m anomalies with the lowest values close to -8°C. Meanwhile, XGBoost partially predicts the correct spatial pattern but with conservative values in the range of $[-1.5°\mathrm{C}, 1.5°\mathrm{C}]$, which are much smaller than the magnitudes of the ground truth. These two examples demonstrate that the SubX models have a certain advantage in matching the amplitude of the tmp2m anomalies, while the ML models are more conservative and provide predicted values with smaller amplitude. On the flip side, in situations where the SubX models do not predict the spatial pattern correctly, the forecasts can be wrong by a large amount.

Acknowledging the advantages of both types of models, we explore a suitable combination of the ML models and the SubX forecasts. More specifically, we investigate whether including SubX forecasts in the feature set of the ML models can enhance the predictive skill of the ML models. Since the hindcast periods of the SubX models are $\sim$10 years shorter than the temporal range of the training data for the ML models, and the temporal resolution of SubX models is also relatively lower, incorporating the SubX forecasts significantly reduces the sample size. To fairly compare the performance, we first train a ML model using the samples that are available during the hindcast periods and then compare it with the ML model that uses SubX forecasts as features, this guarantees both models are trained with exactly the same sample size. For Multitask Lasso, features are originally shared for all locations. To incorporate SubX forecasts, we build one Lasso model for each location but the hyper-parameter is jointly selected based on the performance for all locations.

Temporal results using XGBoost and Lasso, with and without the inclusion of SubX forecasts in the feature set are shown in Figure 6. The detailed spatial results are reported in the Appendix. Overall adding either GMAO-GEOS forecasts in the feature set leads to a significant enhancement of predictive skills. As shown in Figure 6, the combination of the ML models and the SubX forecasts effectively converts some negative temporal ACC to positive and strengthens the forecasts originally achieving positive temporal ACC. The improvement is particularly outstanding for the west-north-central region, a region considered hard to predict. Regarding temporal relative R^2, both ML models obtain some improvements in the areas originally characterized by negative values. Especially for Lasso, it picks the central area where GMAO-GEOS performs well and obtains positive temporal relative R^2. These results highlight the potential to further increase the predictive skill of the ML models by incorporating SubX forecasts. We anticipate that more hindcast data from SubX models would lead to notable improvements in the predictive performance of the ML models.

7 Discussion & Conclusions

In this paper, sub-seasonal climate forecasting, an important but challenging scientific problem, is introduced to the artificial intelligence community. We perform a rigorous evaluation and comparison between state-of-the-art machine learning models and two dynamical models from the SubX project, i.e., GMAO-GEOS and NCEP-CFSv2, for SSF in the western contiguous U.S. Experimental results demonstrate that, on average, the ML models can outperform the SubX models. However, the ML model forecasts usually are relatively conservative compared to the SubX forecasts which, when correctly made, match the scale of the ground truth better. Acknowledging the strengths of both ML and dynamical models, we obtain significant improvements in predictive performance by including the SubX forecasts as a new feature of ML models, which illustrates the potential in generating skillful SSF by combining such two types of models. Further, we show that ML models make most of the bad forecasts during weather extremes, e.g., unusual cold waves, and suggest ways of further improving the ML models by modeling extreme events separately.

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 high-performance computing support from Casper provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation.

References

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.

Butler, A. H.; Seidel, D. J.; Hardiman, S. C.; Butchart, N.; Birner, T.; and Match, A. 2015. Defining Sudden Stratospheric Warmings. *Bulletin of the American Meteorological Society*, 96(11): 1913 – 1928.

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.

DelSole, T.; Trenary, L.; Tippett, M. K.; and Pegion, K. 2017a. Predictability of week-3–4 average temperature and precipitation over the contiguous United States. *Journal of Climate*, 30(10): 3499–3512.

DelSole, T.; Trenary, L.; Tippett, M. K.; and Pegion, K. 2017b. Predictability of Week-3–4 Average Temperature and Precipitation over the Contiguous United States. *Journal of Climate*, 30(10): 3499 – 3512.

Domeisen, D. I.; and Butler, A. H. 2020. Stratospheric drivers of extreme events at the Earth's surface. *Communications Earth & Environment*, 1(1): 1–8.

Dueben, P. D.; and Bauer, P. 2018. Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, 11(10): 3999–4009.

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).

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

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.

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.; DelSole, T.; Ravikumar, P.; and Banerjee, A. 2021. Sub-Seasonal Climate Forecasting via Machine Learning: Challenges, Analysis, and Advances. *AAAI Conference on Artificial Intelligence (AAAI).*

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.

Karl, T.; and Koss, W. J. 1984. Regional and national monthly, seasonal, and annual temperature weighted by area, 1895-1983. *Historical climatology series ; 4-3*.

Koster, R. D.; Mahanama, S. P. P.; Yamada, T. J.; Balsamo, G.; Berg, A. A.; Boisserie, M.; Dirmeyer, P. A.; Doblas-Reyes, F. J.; Drewitt, G.; Gordon, C. T.; Guo, Z.; Jeong, J.-H.; Lee, W.-S.; Li, Z.; Luo, L.; Malyshev, S.; Merryfield, W. J.; Seneviratne, S. I.; Stanelle, T.; van den Hurk, B. J. J. M.; Vitart, F.; and Wood, E. F. 2011. The Second Phase of the Global Land–Atmosphere Coupling Experiment: Soil Moisture Contributions to Subseasonal Forecast Skill. *Journal of Hydrometeorology*, 12(5): 805 – 822.

Li, S.; and Robertson, A. W. 2015. Evaluation of submonthly precipitation forecast skill from global ensemble prediction systems. *Monthly Weather Review*, 143(7): 2871–2889.

Liu, Y.; Racah, E.; Correa, J.; Khosrowshahi, A.; Lavers, D.; Kunkel, K.; Wehner, M.; Collins, W.; et al. 2016. Application of deep convolutional neural networks for detecting extreme weather in climate datasets. *arXiv preprint arXiv:1605.01156*.

Lorenz, E. N. 1963. Deterministic nonperiodic flow. *Journal of atmospheric sciences*, 20(2): 130–141.

Mariotti, A.; Baggett, C.; Barnes, E. A.; Becker, E.; Butler, A.; Collins, D. C.; Dirmeyer, P. A.; Ferranti, L.; Johnson, N. C.; Jones, J.; et al. 2020. Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bulletin of the American Meteorological Society*, 101(5): E608–E625.

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.

NIDIS. 2019. Forecast Rodeo II Leaderboard. https://www.drought.gov/forecast-rodeo-ii-leaderboard. Accessed: 2021-09-08.

Pegion, K.; Kirtman, B. P.; Becker, E.; Collins, D. C.; LaJoie, E.; Burgman, R.; Bell, R.; DelSole, T.; Min, D.; Zhu, Y.; Li, W.; Sinsky, E.; Guan, H.; Gottschalck, J.; Metzger, E. J.; Barton, N. P.; Achuthavarier, D.; Marshak, J.; Koster, R. D.; Lin, H.; Gagnon, N.; Bell, M.; Tippett, M. K.; Robertson, A. W.; Sun, S.; Benjamin, S. G.; Green, B. W.; Bleck, R.; and Kim, H. 2019. The Subseasonal Experiment (SubX): A Multimodel Subseasonal Prediction Experiment. *Bulletin of the American Meteorological Society*, 100(10): 2043–2060.

Reichle, R. H.; and Liu, Q. 2014. Observation-corrected precipitation estimates in GEOS-5. *Technical Report Series on Global Modeling and Data Assimilation*, 35.

Saha, S.; Moorthi, S.; Wu, X.; Wang, J.; Nadiga, S.; Tripp, P.; Behringer, D.; Hou, Y.-T.; Chuang, H.-y.; Iredell, M.; et al. 2014. The NCEP climate forecast system version 2. *Journal of climate*, 27(6): 2185–2208.

Scher, S.; and Messori, G. 2019. Weather and climate forecasting with neural networks: using general circulation models (GCMs) with different complexity as a study ground. *Geoscientific Model Development*, 12(7): 2797–2809.

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.

Srinivasan, V.; Khim, J.; Banerjee, A.; and Ravikumar, P. 2021. Subseasonal Climate Prediction in the Western US using Bayesian Spatial Models. *Conference on Uncertainty in Artificial Intelligence (UAI)*.

Stan, C.; Straus, D. M.; Frederiksen, J. S.; Lin, H.; Maloney, E. D.; and Schumacher, C. 2017. Review of Tropical-Extratropical Teleconnections on Intraseasonal Time Scales. *Reviews of Geophysics*, 55(4): 902–937.

Stevens, A.; Willett, R.; Mamalakis, A.; Foufoula-Georgiou, E.; Tejedor, A.; Randerson, J. T.; Smyth, P.; and Wright, S. 2021. Graph-guided regularized regression of Pacific Ocean climate variables to increase predictive skill of southwestern US winter precipitation. *Journal of Climate*, 34(2): 737–754.

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

Trewin, B.; Baddour, O.; Organization, W. M.; Kontongomde, H.; Data, W. C.; and Programme, M. 2007. *The Role of Climatological Normals in a Changing Climate*. WCDMP (Series). World Meteorological Organization.

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

USBR and NOAA. 2019. Forecast Rodeo II. https://www.challenge.gov/challenge/rodeo-ii-sub-seasonal-climate-forecasting/. Accessed: 2021-09-08.

Vitart, F. 2004. Monthly forecasting at ECMWF. *Monthly Weather Review*, 132(12): 2761–2779.

Vitart, F. 2014. Evolution of ECMWF sub-seasonal forecast skill scores. *Quarterly Journal of the Royal Meteorological Society*, 140(683): 1889–1899.

Vitart, F.; Ardilouze, C.; Bonet, A.; Brookshaw, A.; Chen, M.; Codorean, C.; Déqué, M.; Ferranti, L.; Fucile, E.; Fuentes, M.; et al. 2017. The subseasonal to seasonal (S2S) prediction project database. *Bulletin of the American Meteorological Society*, 98(1): 163–173.

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).

Waliser, D. 2005. *Predictability and forecasting*, 389–423. Berlin, Heidelberg: Springer Berlin Heidelberg.

Weigel, A. P.; Baggenstos, D.; Liniger, M. A.; Vitart, F.; and Appenzeller, C. 2008. Probabilistic verification of monthly temperature forecasts. *Monthly Weather Review*, 136(12): 5162–5182.

Weyn, J. A.; Durran, D. R.; Caruana, R.; and Cresswell-Clay, N. 2021. Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. *arXiv preprint arXiv:2102.05107*.

White, C. J.; Carlsen, H.; Robertson, A. W.; Klein, R. J.; Lazo, J. K.; Kumar, A.; Vitart, F.; Coughlan de Perez, E.; Ray, A. J.; Murray, V.; et al. 2017. Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteorological applications*, 24(3): 315–325.

Wikipedia. 2019. January–February 2019 North American cold wave. https://en.wikipedia.org/wiki/January-February_2019_North_American_cold_wave. Accessed: 2021-09-08.