Machine Learning and Dynamical Models for Sub-seasonal Climate Forecasting

Sijie He
Dept. of Computer Science & Engineering
University of Minnesota, Twin Cities
Minneapolis, MN 55455
hexxx893@umn.edu

Xinyan Li
Dept. of Computer Science & Engineering
University of Minnesota, Twin Cities
Minneapolis, MN 55455
lixx1166@umn.edu

Laurie Trenary
Dept. of Atmospheric, Oceanic and Earth Sciences
George Mason University
Fairfax, VA 22030
ltrenary@gmu.edu

Benjamin A Cash
Dept. of Atmospheric, Oceanic and Earth Sciences
George Mason University
Fairfax, VA 22030
bcash@gmu.edu

Timothy DelSole
Dept. of Atmospheric, Oceanic and Earth Sciences
George Mason University
Fairfax, VA 22030
tdelsole@gmu.edu

Arindam Banerjee
Dept. of Computer Science
University of Illinois Urbana-Champaign
Urbana, IL 61801
arindamb@illinois.edu

Abstract

Sub-seasonal climate forecasting (SSF) is the prediction of key climate variables such as temperature and precipitation on a 2-week to 2-month time horizon. Skillful SSF would have substantial societal value in areas such as agricultural productivity, water resource management, and emergency planning for droughts and wildfires. Despite its societal importance, SSF has stayed a challenging problem and mainly relies on physics-based dynamical models. Meanwhile, recent studies have shown the potential of machine learning (ML) models to advance SSF. In this paper, we show that suitably incorporating dynamical model forecasts as inputs to ML models can substantially improve their forecasting performance. The SSF dataset constructed for the work, dynamical model predictions, and code for the ML models are released along with the paper for the benefit of the broader machine learning community [1].

1 Introduction

Over the past decade, good quality short-term (few days) weather forecasts as well as 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

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

Fourth Workshop on Machine Learning and the Physical Sciences (NeurIPS 2021).

not yet available. Skillful SSF has immense societal value as discussed in two recent reports from the National Academy of Sciences (NAS) [12, 11]. 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 [24].

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 [21]. Due to the chaotic nature of the atmosphere, weather events can not be accurately predicted beyond two weeks using dynamical models [10]. 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 an unconventional time series prediction problem. 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 small τ_s. 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 along with the nonlinear dynamics and complex interactions makes SSF challenging.

To understand the conditions that lead to enhanced predictability and to improve sub-seasonal to seasonal (S2S) forecasts, some research-to-operations projects such as S2S [20] and Subseasonal Experiment (SubX) [13] have been established. Currently, sub-seasonal forecasts based on dynamical models are available weekly through the SubX project, which contains 7 different global prediction models contributing predictions over the period 1999 to present [13]. The project provides real-time and retrospective forecasts for community exploration of sub-seasonal prediction, but the full utility of the forecasts for operational forecasting still remains to be determined. Meanwhile, ML models have started to be explored for predictions of temperature, precipitation, and other climate variables on sub-seasonal time scales [7, 6, 22, 17], which have shown great promise for SSF. In this paper, we consider enhancing ML models by using forecasts from state-of-the-art dynamical models from the SubX project. More specifically, we include physics-based dynamical model forecasts as covariates in the ML models. The empirical results illustrate that using dynamical model forecasts as inputs improves the ML model forecasts, and the improvements are statistically significant. In addition, we release all the data, as well as code to replicate and hopefully extend our work.

2 Sub-seasonal Climate Forecasting

Problem Statement. In climate science, it is more important to predict temperature anomalies than absolute temperatures. A temperature anomaly is the difference from a baseline temperature, e.g., climatology, which is typically a historical long-term average temperature for each calendar day at each geographic location. A positive temperature anomaly indicates the temperature is warmer than the historical average, vice versa. In this paper, 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. 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 is from 2017 to 2019.

Ground Truth Dataset. The ground truth dataset is constructed from NOAA's Climate Prediction Center (CPC) Global Gridded Temperature dataset [4], which is commonly applied for forecast verification by NOAA/CPC [4]. The CPC dataset provides daily max and min 2m temperatures (tmp2m), which refers to air temperature at 2 meters above the surface, from Jan 1, 1979, to present. The daily temperature anomalies are computed by subtracting the climatology from the observed daily tmp2m, where the climatology is the smoothed long-term average of tmp2m over 1990 - 2016 for each month-day combination and grid point. The forecasting target at each date and grid point is the average of tmp2m anomalies at days 15 to 28 (weeks 3 & 4).

3 Machine Learning and Dynamical Models

Machine Learning Models. In this paper, we focus on two machine learning models which have been shown to work effectively for sub-seasonal climate forecasting [6].

Gradient boosted trees (XGBoost) [1]. A functional gradient boosting algorithm, of which the weak learners are regression trees. The algorithm combines multiple weak learners into a stronger 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 best hyper-parameters, such as the maximum depth of a tree, number of gradient boosted trees, and learning rate, are selected jointly based on the performance over all locations.

Lasso [18, 8]. A regularized linear regression model. Denote $y_{g,t} \in \mathbb{R}$ and $\mathbf{x}_{g,t} \in \mathbb{R}^p$ as the response and covariates for a date t and a location g. We assume $y_{g,t} = \mathbf{x}_{g,t}^T \theta_g^* + \epsilon$, where $\epsilon \in \mathbb{R}$ is a Gaussian noise and $\theta_g^* \in \mathbb{R}^p$ is the coefficient for location g. The coefficient θ_g^* is estimated by $\hat{\theta}_g = \operatorname{argmin}_{\theta_g \in \mathbb{R}^p} \frac{1}{2T} \|Y_g - X_g \theta_g\|_2^2 + \lambda \|\theta_g\|_1$, where $Y_g \in \mathbb{R}^T$ and $X_g \in \mathbb{R}^{T \times p}$ are the tmp2m anomalies and covariates for the location g over T dates, respectively. λ is the penalty parameter shared by all locations.

Covariates. We consider two types of covariates for the ML models, i.e., observational climate data and forecasts from SubX models.

Observational Climate Data. We select a suitable suite of climate variables representing the condition of atmosphere, land, and ocean. Spatially over the contiguous U.S., we consider 2m temperature [4], soil moisture [3], geopotential height, sea level pressure and relative humidity [9]. We also obtain sea surface temperature [15] 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. 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. The extracted PCs are then normalized by z-scoring for each month-day combination separately. In addition, we consider climate indices that describe the state of the climate system or are related to different climate phenomena, including Multivariate ENSO index [27], Niño indices [15], North Atlantic Oscillation index [19], Madden-Julian Oscillation indices [23], and Sudden Stratospheric Warming index [5]. The climate indices and the PC-based features of all spatiotemporal climate variables jointly form the feature set for each date.

SubX Forecasts. We focus on one SubX model, the Climate Forecast System version 2 (CFSv2) from National Centers for Environmental Prediction (NCEP) [16]. NCEP-CFSv2 is a coupled atmosphere–ocean–land–ice model and is the operational seasonal prediction model used by the U.S. Climate Prediction Center. The SubX model has two predictive periods: hindcast and forecast. A hindcast period (1999-2015) represents the time when a dynamic model re-forecasts historical events and a forecast period (from July 2017) has real-time predictions generated daily. NCEP-CFSv2 includes four ensemble members and the average of four ensemble members' outputs are taken as the forecasts. All forecasts include daily values for 45 days beyond the initialization date. The weeks 3 & 4 outlooks are computed by averaging the forecasts at days 15 - 28 ahead and subtracting the corresponding climatology computed from the hindcast period.

4 Experimental Setup and Results

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 and separate predictive models are trained accordingly. The best hyper-parameters of each type of ML model are selected on a monthly basis. To do so, for each month of the year, e.g. January, we construct five validation sets containing data from the same month (e.g., January) in 2011 - 2015, and the corresponding training sets consist of 12 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 18 years of data prior to a given test set to train the corresponding forecasting model.

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

Cosine Similarity 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. Cosine similarity, also known as *uncentered anomaly correlation* [26], is the only metric used in the Sub-Seasonal Climate Forecast Rodeo Competition [14, 7].

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.

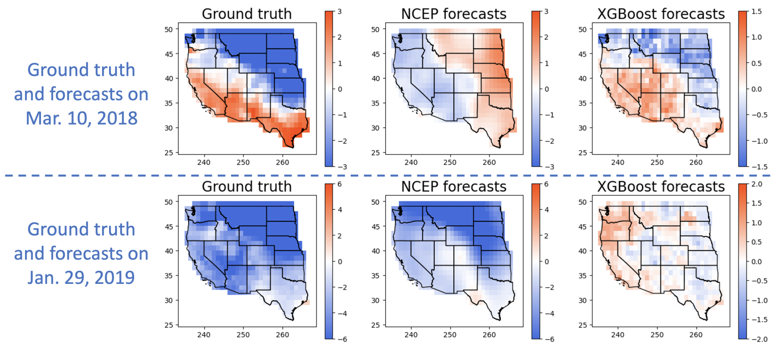


Figure 1: Comparison between the ground truth and forecasts made by NCEP-CFSv2 and XGBoost on March 10, 2018 (top), which is an example when SubX forecasts mistakenly predict the pattern with a large magnitude, and Jan 29, 2019 (bottom), when SubX forecasts predict the extreme weather well. In both cases, predicted values from XGBoost are much smaller than NCEP-CFSv2 forecasts.

Denote the ground truth temperature anomalies as $Y^* \in \mathbb{R}^{T \times G}$ for T dates and G grid points. The *spatial* predictive skill for a given date t can be evaluated on $Y^*[t, :]$ (the t-th row of Y^*), which corresponds to the ground truth for all grid points at t. The *temporal* predictive skill for a grid point g can be evaluated on $Y^*[:, g]$ (the g-th column of Y^*), similar to time series prediction evaluation.

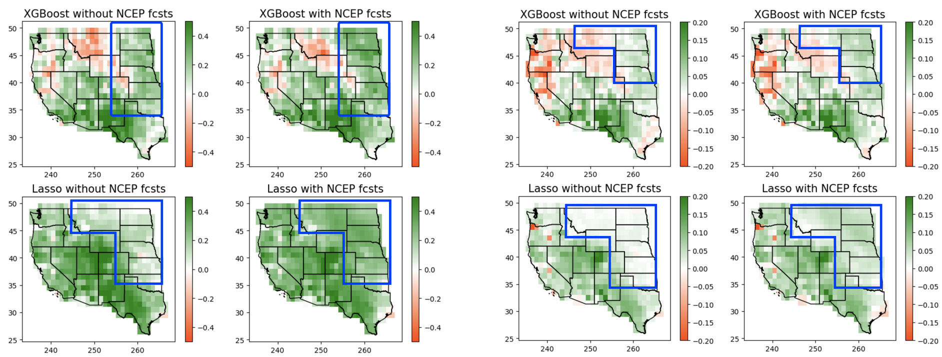
Experimental Results. To demonstrate the strengths and limitations of the SubX and the ML model forecasts, we present forecasts of two days as anecdotal evidence in Figure 1. The first example shows that, on Mar 10, 2018, XGBoost has reproduced the spatial pattern of the ground truth, while NCEP-CFSv2 predicts an opposite pattern. The second example on Jan 29, 2019, illustrates that, when a cold wave affected the U.S. leading to extreme low average tmp2m anomalies [25], the SubX forecasts successfully estimate the spatial pattern of the ground truth, while XGBoost partially predicts the spatial pattern. In both examples, the predicted scale from NCEP-CFSv2 is much larger than XGBoost, and is closer to the scale of the ground truth. These two examples demonstrate that the SubX models have certain advantage on matching the magnitude of the tmp2m anomalies, while the ML models are more conservative. 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 their predictive skill. To compare the performance fairly, 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.

Table 1: The mean and median (standard error) of spatial cosine similarity and spatial relative R^2 of XGBoost and Lasso with and without including the SubX forecasts in their feature set.

Methods	Features	cos w/o NCEP	cos with NCEP	relative R^2 w/o NCEP	relative R^2 with NCEP
XGBoost	Mean	0.15 (0.02)	**0.18** (0.02)	0.03 (0.01)	**0.04** (0.01)
	Median	0.21 (0.03)	**0.23** (0.02)	0.04 (0.01)	**0.04** (0.01)
Lasso	Mean	0.19 (0.01)	**0.23** (0.02)	0.03 (0.00)	**0.05** (0.01)
	Median	0.21 (0.02)	**0.25** (0.02)	0.03 (0.00)	**0.04** (0.01)

Table 1 presents the spatial cosine similarity and relative R^2 using XGBoost and Lasso, with and without the inclusion of SubX forecasts in the feature set. The temporal results are shown in Figure 2. Adding NCEP-CFSv2 forecasts in the feature set leads to a significant enhancement of predictive skill. We conduct the sign test introduced in [2] to compare the differences in forecast skills. Overall, comparison of ML models' performance with and without SubX features yields p values much smaller than 0.01. Furthermore, as shown in Figure 2, the combination of the ML models and the SubX forecasts effectively converts some negative temporal cosine similarity to positive and strengthens the forecasts originally achieving positive temporal cosine similarity. The improvement



(a) Temporal cosine similarity (b) Temporal relative R^2

Figure 2: The temporal cosine similarity and relative R^2 of XGBoost and Lasso with and without NCEP-CFSv2 forecasts as features. Including SubX forecasts in the feature set evidently improves the forecasting performance, especially for the central U.S. (top right corner, marked by blue frames).

is particularly outstanding for the west-north-central U.S., where temperature fluctuations are more drastic compared to the coastal areas. Similarly, regarding temporal relative R^2, both ML models obtain some improvements in the areas originally characterized by values close to 0 (white). These results highlight the potential to further increase predictive skill of the ML models by incorporating SubX forecasts.

However, to include SubX forecasts in the feature set, the size of the training set is decided by the availability of the hindcast period for SubX models. For example, the features constructed from observational climate variables are available from 1990, while the hindcast period of NCEP-CFSv2 starts in 1999. The relatively short hindcast period leads to a reduction of almost one third of the data for model training. The small sample size restrains the forecasting performance of the XGBoost model, which explains why Lasso achieves higher prediction accuracy than XGBoost. We anticipate that more hindcast data from SubX models would lead to a notable improvement in the predictive skills of the ML models.

5 Discussion & Conclusions

In this paper, sub-seasonal climate forecasting, an important but challenging scientific problem, is introduced to the machine learning community. Acknowledging the strengths of both machine learning and physics-based dynamical models, we explore the potential in generating skillful SSF by combining such two types of models. We obtain significant improvements in predictive performance by including the SubX forecasts as a new feature of ML models. To extend our work, which focuses on ML models for deterministic forecasting in SSF, future studies could explore probabilistic ML models to obtain the uncertainty of the forecasts.

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

[1] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 785–794, 2016.

[2] Timothy DelSole and Michael K Tippett. Forecast comparison based on random walks. *Monthly Weather Review*, 144(2):615–626, 2016.

[3] Yun Fan and Huug van den Dool. Climate prediction center global monthly soil moisture data set at 0.5 resolution for 1948 to present. *Journal of Geophysical Research: Atmospheres*, 109(D10), 2004.

[4] Yun Fan and Huug Van den Dool. A global monthly land surface air temperature analysis for 1948–present. *Journal of Geophysical Research: Atmospheres*, 113(D1), 2008.

[5] Ronald Gelaro, Will McCarty, Max J Suárez, Ricardo Todling, Andrea Molod, Lawrence Takacs, Cynthia A Randles, Anton Darmenov, Michael G Bosilovich, Rolf Reichle, et al. The modern-era retrospective analysis for research and applications, version 2 (merra-2). *Journal of climate*, 30(14):5419–5454, 2017.

[6] Sijie He, Xinyan Li, Timothy DelSole, Pradeep Ravikumar, and Arindam Banerjee. Sub-seasonal climate forecasting via machine learning: Challenges, analysis, and advances. *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[7] Jessica Hwang, Paulo Orenstein, Judah Cohen, Karl Pfeiffer, and Lester Mackey. Improving subseasonal forecasting in the western us with machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2325–2335. ACM, 2019.

[8] Ali Jalali, Pradeep Ravikumar, and Sujay Sanghavi. A dirty model for multiple sparse regression. *IEEE Transactions on Information Theory*, 59(12):7947–7968, 2013.

[9] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, Roy Jenne, and Dennis Joseph. The ncep/ncar 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77(3):437–472, 1996.

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

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

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

[13] Kathy Pegion, Ben P. Kirtman, Emily Becker, Dan C. Collins, Emerson LaJoie, Robert Burgman, Ray Bell, Timothy DelSole, Dughong Min, Yuejian Zhu, Wei Li, Eric Sinsky, Hong Guan, Jon Gottschalck, E. Joseph Metzger, Neil P Barton, Deepthi Achuthavarier, Jelena Marshak, Randal D. Koster, Hai Lin, Normand Gagnon, Michael Bell, Michael K. Tippett, Andrew W. Robertson, Shan Sun, Stanley G. Benjamin, Benjamin W. Green, Rainer Bleck, and Hyemi Kim. The subseasonal experiment (subx): A multimodel subseasonal prediction experiment. *Bulletin of the American Meteorological Society*, 100(10):2043–2060, 2019.

[14] David Raff, Kenneth Nowak, Robert Cifelli, Levi D Brekke, and Robert Stabler Webb. Sub-seasonal climate forecast rodeo. In *AGU Fall Meeting*, 2017.

[15] Richard W Reynolds, Thomas M Smith, Chunying Liu, Dudley B Chelton, Kenneth S Casey, and Michael G Schlax. Daily high-resolution-blended analyses for sea surface temperature. *Journal of Climate*, 20(22):5473–5496, 2007.

[16] Suranjana Saha, Shrinivas Moorthi, Xingren Wu, Jiande Wang, Sudhir Nadiga, Patrick Tripp, David Behringer, Yu-Tai Hou, Hui-ya Chuang, Mark Iredell, et al. The ncep climate forecast system version 2. *Journal of climate*, 27(6):2185–2208, 2014.

[17] Vishwak Srinivasan, Justin Khim, Arindam Banerjee, and Pradeep Ravikumar. Subseasonal climate prediction in the western us using bayesian spatial models. *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2021.

[18] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society,*, pages 267–288, 1996.

[19] HM Van den Dool, S Saha, and AAke Johansson. Empirical orthogonal teleconnections. *Journal of Climate*, 13(8):1421–1435, 2000.

[20] F Vitart, C Ardilouze, A Bonet, A Brookshaw, M Chen, C Codorean, M Déqué, L Ferranti, E Fucile, M Fuentes, et al. The subseasonal to seasonal (s2s) prediction project database. *Bulletin of the American Meteorological Society*, 98(1):163–173, 2017.

[21] Frédéric Vitart, Andrew W Robertson, and David LT Anderson. Subseasonal to seasonal prediction project: Bridging the gap between weather and climate. *Bulletin of the World Meteorological Organization*, 61(23), 2012.

[22] Jonathan A Weyn, Dale R Durran, Rich Caruana, and Nathaniel Cresswell-Clay. Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. *arXiv preprint arXiv:2102.05107*, 2021.

[23] Matthew C. Wheeler and Harry H. Hendon. An all-season real-time multivariate mjo index: Development of an index for monitoring and prediction. *Monthly Weather Review*, 132(8):1917–1932, 2004.

[24] Christopher J White, Henrik Carlsen, Andrew W Robertson, Richard JT Klein, Jeffrey K Lazo, Arun Kumar, Frederic Vitart, Erin Coughlan de Perez, Andrea J Ray, Virginia Murray, et al. Potential applications of subseasonal-to-seasonal (s2s) predictions. *Meteorological applications*, 24(3):315–325, 2017.

[25] Wikipedia. January–February 2019 North American cold wave. `https://en.wikipedia.org/wiki/January-February_2019_North_American_cold_wave`, 2019.

[26] Daniel S Wilks. *Statistical methods in the atmospheric sciences*, volume 100. Academic press, 2011.

[27] Tao Zhang, Andrew Hoell, Judith Perlwitz, Jon Eischeid, Donald Murray, Martin Hoerling, and Thomas M Hamill. Towards probabilistic multivariate enso monitoring. *Geophysical Research Letters*, 46(17-18):10532–10540, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]