Improving Subseasonal Forecasting in the Western U.S. with Machine Learning

Jessica Hwang
Stanford University
jjhwang@stanford.edu

Paulo Orenstein
Stanford University
pauloo@stanford.edu

Karl Pfeiffer
Atmospheric and Environmental Research
kpfeiffe@aer.com

Judah Cohen
Atmospheric and Environmental Research
jcohen@aer.com

Lester Mackey
Microsoft Research New England
lmackey@microsoft.com

Abstract

Water managers in the western United States (U.S.) rely on longterm forecasts of temperature and precipitation to prepare for droughts and other wet weather extremes. To improve the accuracy of these longterm forecasts, the Bureau of Reclamation and the National Oceanic and Atmospheric Administration (NOAA) launched the Subseasonal Climate Forecast Rodeo, a year-long real-time forecasting challenge, in which participants aimed to skillfully predict temperature and precipitation in the western U.S. two to four weeks and four to six weeks in advance. We present and evaluate our machine learning approach to the Rodeo and release our `SubseasonalRodeo` dataset, collected to train and evaluate our forecasting system. Our system is an ensemble of two regression models, and exceeds that of the top Rodeo competitor as well as the government baselines for each target variable and forecast horizon. The full paper is Hwang et al. (2018b).

1 Main Results

Water and fire managers in the western United States (U.S.) rely on *subseasonal forecasts*—forecasts of temperature and precipitation two to six weeks in advance—to allocate water resources, manage wildfires, and prepare for droughts and other weather extremes (White et al., 2017). While purely physics-based numerical weather prediction dominates the landscape of short-term weather forecasting, such deterministic methods have a limited *skillful* (i.e., accurate) forecast horizon due to the chaotic nature of weather (Lorenz, 1963). Prior to the widespread availability of operational numerical weather prediction, weather forecasters made predictions using their knowledge of past weather patterns and climate (sometimes called *the method of analogs*) (Nebeker, 1995). The current availability of ample meteorological records and high-performance computing offers the opportunity to blend physics-based and statistical machine learning (ML) approaches to extend the skillful forecast horizon.

This data and computing opportunity, coupled with the critical operational need, motivated the U.S. Bureau of Reclamation and the National Oceanic and Atmospheric Administration (NOAA) to conduct the Subseasonal Climate Forecast Rodeo (Nowak et al., 2017), a year-long real-time forecasting challenge, in which participants aimed to skillfully predict temperature and precipitation

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

in the western U.S. two to four weeks and four to six weeks in advance. To meet this challenge, we developed an ML-based forecasting system and a `SubseasonalRodeo` dataset (Hwang et al., 2018a) suitable for training and benchmarking subseasonal forecasts.

Our subseasonal ML system is an ensemble of two regression models: a local linear regression model with multitask model selection (`MultiLLR`) and a weighted local autoregression enhanced with multitask k-nearest neighbor features (`AutoKNN`). The `MultiLLR` model introduces candidate regressors from each data source in the `SubseasonalRodeo` dataset and then prunes irrelevant predictors using a multitask backward stepwise criterion designed for the forecasting skill objective. The `AutoKNN` model extracts features only from the target variable (temperature or precipitation), combining lagged measurements with a skill-specific form of nearest-neighbor modeling. For each of the two Rodeo target variables (temperature and precipitation) and forecast horizons (weeks 3-4 and weeks 5-6), our work makes the following principal contributions:

1. We release a new `SubseasonalRodeo` dataset suitable for training and benchmarking subseasonal forecasts.
2. We introduce two subseasonal regression approaches tailored to the forecast skill objective, one of which uses only features of the target variable.
3. We introduce a simple ensembling procedure that provably improves average skill whenever average skill is positive.
4. We show that each regression method alone outperforms the Rodeo benchmarks, including a debiased version of the operational U.S. Climate Forecasting System (CFSv2), and that our ensemble outperforms the top Rodeo competitor.
5. We show that, over 2011-2018, an ensemble of our models and debiased CFSv2 improves debiased CFSv2 skill by 37-53% for temperature and 128-154% for precipitation.

For details, see the full paper (Hwang et al., 2018b).

2 Forecasting Challenge Details

The Subseasonal Climate Forecast Rodeo was a year-long, real-time forecasting competition in which, every two weeks, contestants submitted forecasts for average temperature (°C) and total precipitation (mm) at two forecast horizons, 15-28 days ahead (weeks 3-4) and 29-42 days ahead (weeks 5-6). The geographic region of interest was the western contiguous United States, delimited by latitudes 25N to 50N and longitudes 125W to 93W, at a 1° by 1° resolution, for a total of $G = 514$ grid points. The initial forecasts were issued on April 18, 2017 and the final on April 3, 2018.

Forecasts were judged on the spatial cosine similarity between predictions and observations adjusted by a long-term average. More precisely, let t denote a date represented by the number of days since January 1, 1901, and let $\texttt{year}(t)$, $\texttt{doy}(t)$, and $\texttt{monthday}(t)$ respectively denote the year, the day of the year, and the month-day combination (e.g., January 1) associated with that date. We associate with the two-week period beginning on t an observed average temperature or total precipitation $\mathbf{y}_t \in \mathbf{R}^G$ and an observed *anomaly*

$$\mathbf{a}_t = \mathbf{y}_t - \mathbf{c}_{\texttt{monthday}(t)}, \tag{1}$$

where

$$\mathbf{c}_d \triangleq \tfrac{1}{30} \sum_{\substack{t\,:\,\texttt{monthday}(t)=d,\\ 1981\leq\texttt{year}(t)\leq 2010}} \mathbf{y}_t \tag{2}$$

is the *climatology* or long-term average over 1981-2010 for the month-day combination d. Contestant forecasts $\hat{\mathbf{y}}_t$ were judged on the cosine similarity—termed *skill* in meteorology—between their forecast anomalies $\hat{\mathbf{a}}_t = \hat{\mathbf{y}}_t - \mathbf{c}_{\texttt{monthday}(t)}$ and the observed anomalies:

$$\text{skill}(\hat{\mathbf{a}}_t, \mathbf{a}_t) \triangleq \cos(\hat{\mathbf{a}}_t, \mathbf{a}_t) = \tfrac{\langle \hat{\mathbf{a}}_t, \mathbf{a}_t \rangle}{\|\hat{\mathbf{a}}_t\|_2 \|\mathbf{a}_t\|_2}. \tag{3}$$

To qualify for a prize, contestants had to achieve higher mean skill over all forecasts than two government benchmarks, a debiased version of the physics-based operational U.S. Climate Forecasting System (CFSv2) and a damped persistence forecast. The official contest CFSv2 forecast for t, an average of 32 operational forecasts based on 4 model initializations and 8 lead times, was debiased by adding the mean observed temperature or precipitation for $\texttt{monthday}(t)$ over 1999-2010 and subtracting the mean CFSv2 reforecast, an average of 8 lead times for a single model initialization, over the same period. An exact description of the damped persistence model was not provided.

Table 1: Average contest-period skill of the proposed models `MultiLLR` and `AutoKNN`, the proposed ensemble of `MultiLLR` and `AutoKNN` (*ensemble*), the official contest debiased-CFSv2 baseline (*cfsv2*), the official contest damped-persistence baseline (*damped*), and the top-performing competitor in the Forecast Rodeo contest (*top competitor*).

task	multillr	autoknn	ensemble	cfsv2	damped	top competitor
temperature, weeks 3-4	0.2856	0.2807	**0.3414**	0.1589	0.1952	0.2855
temperature, weeks 5-6	0.2371	0.2817	**0.3077**	0.2192	-0.0762	0.2357
precipitation, weeks 3-4	0.1675	0.2156	**0.2388**	0.0713	-0.1463	0.2144
precipitation, weeks 5-6	0.2219	0.1870	**0.2412**	0.0227	-0.1613	0.2162

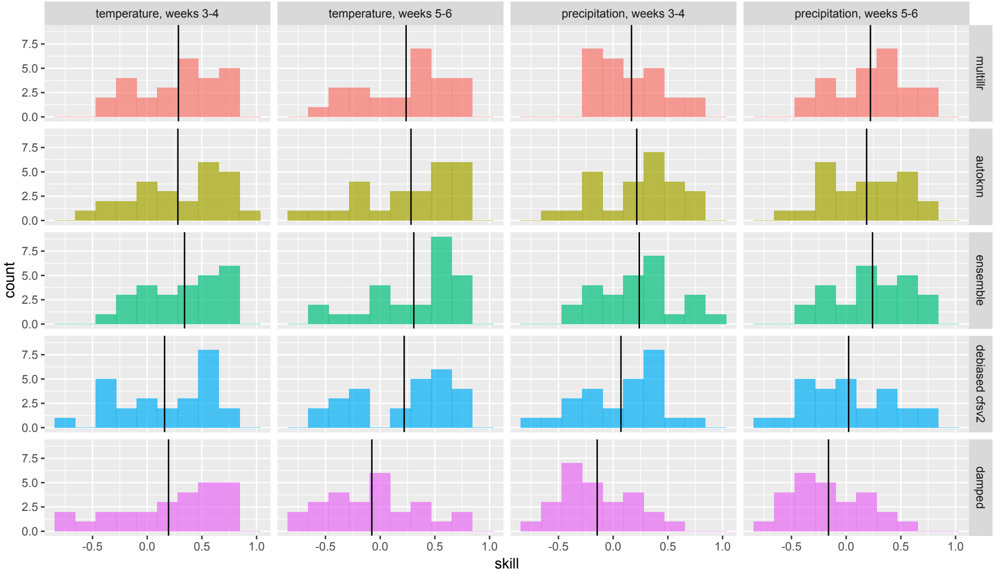


Figure 1: Distribution of contest-period skills of the proposed models `MultiLLR` and `AutoKNN`, the proposed ensemble of `MultiLLR` and `AutoKNN` (*ensemble*), the official contest debiased-CFSv2 baseline, and the official contest damped-persistence baseline (*damped*). Average contest-period skill is indicated by a vertical line.

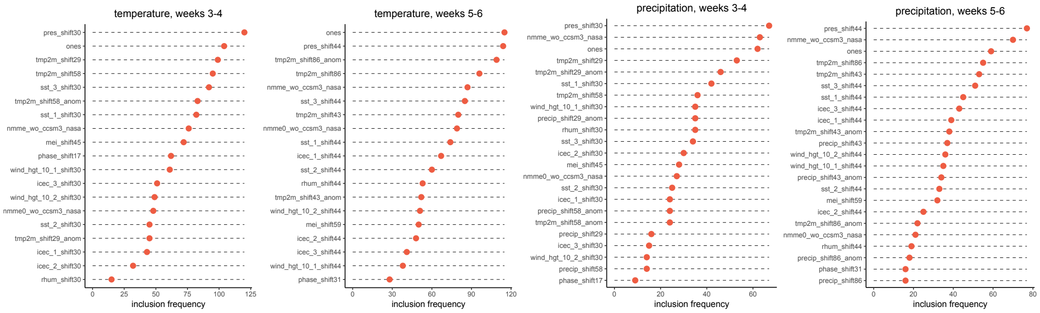
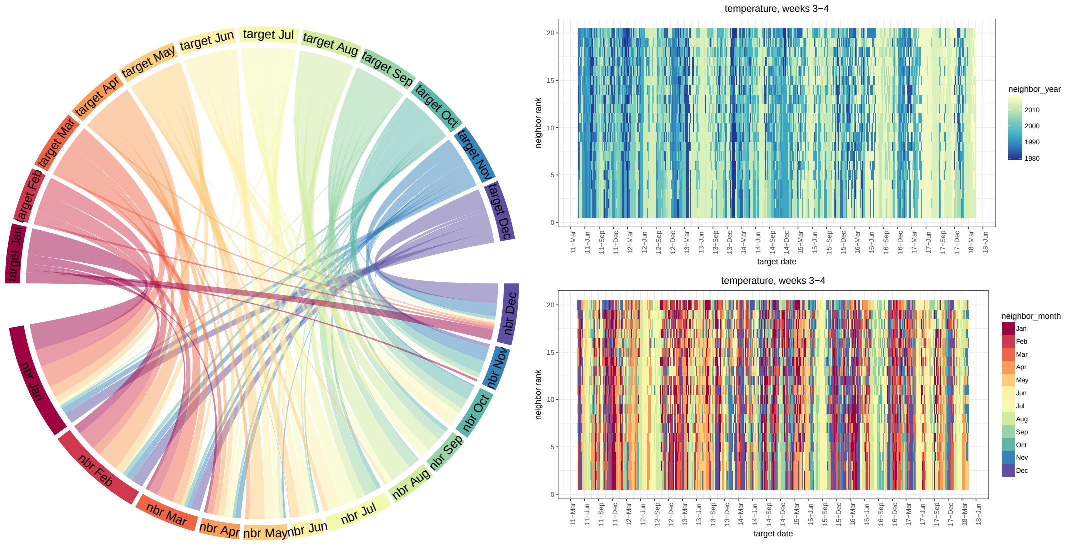


Figure 2: Feature inclusion frequencies of all candidate variables for local linear regression with multitask model selection (`MultiLLR`) across all target dates in the historical forecast evaluation period. For a full description of the variables see Hwang et al. (2018b).

Table 2: Average skills for historical forecasts in each year following the climatology period. We compare the proposed ensemble of the `MultiLLR` and `AutoKNN` models (*ensemble*), the reconstructed debiased CFSv2 baseline (*rec-deb-cfs*), and the proposed ensemble of `MultiLLR`, `AutoKNN`, and debiased CFSv2 (*ens-cfs*).

	temperature, weeks 3-4			temperature, weeks 5-6		
year	ensemble	rec-deb-cfs	ens-cfs	ensemble	rec-deb-cfs	ens-cfs
2011	0.3433	**0.4598**	0.4563	0.3646	0.3879	**0.4405**
2012	0.2173	0.1397	**0.2181**	**0.3529**	0.1030	0.3316
2013	0.1688	**0.2861**	0.2711	**0.1895**	0.1211	0.1858
2014	0.2803	0.3018	**0.3591**	0.2596	0.1936	**0.3311**
2015	0.4339	0.2857	**0.4383**	0.2970	0.4234	**0.4311**
2016	0.3663	0.2490	**0.3887**	**0.3023**	0.0983	0.2799
2017	**0.3414**	0.0676	0.3239	**0.3077**	0.1708	0.2993
all	0.3073	0.2557	**0.3508**	0.2962	0.2142	**0.3279**
	precipitation, weeks 3-4			precipitation, weeks 5-6		
year	ensemble	rec-deb-cfs	ens-cfs	ensemble	rec-deb-cfs	ens-cfs
2011	0.2081	0.1646	**0.2435**	0.2195	0.1835	**0.2704**
2012	**0.3999**	0.0828	0.3854	0.4026	0.1941	**0.4083**
2013	**0.2353**	0.0648	0.1967	**0.1969**	0.0782	0.1915
2014	0.1378	0.1272	**0.1716**	0.0372	0.0155	**0.0537**
2015	0.0396	0.0837	**0.1035**	0.0822	0.0292	**0.0878**
2016	**0.0660**	0.0190	0.0467	**0.0125**	-0.0160	0.0180
2017	**0.2388**	0.0596	0.2270	**0.2412**	-0.0038	0.2026
all	0.1893	0.0860	**0.1964**	0.1703	0.0691	**0.1755**



(a) (b)

Figure 3: (a) Precipitation, weeks 3-4: Distribution of the month of the most similar neighbor learned by `AutoKNN` as a function of the month of the target date. (b) Temperature, weeks 3-4: Year (top) and month (bottom) of the 20 most similar neighbors learned by `AutoKNN` (vertical axis ranges from $k = 1$ to 20) as a function of the target date (horizontal axis).

References

Hwang, J.; Orenstein, P.; Cohen, J.; and Mackey, L. 2018a. The SubseasonalRodeo dataset. *Harvard Dataverse*. `https://doi.org/10.7910/DVN/IHBANG`.

Hwang, J.; Orenstein, P.; Pfeiffer, K.; Cohen, J.; and Mackey, L. 2018b. Improving subseasonal forecasting in the western us with machine learning. *arXiv preprint arXiv:1809.07394*.

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

Nebeker, F. 1995. *Calculating the weather: Meteorology in the 20th century*, volume 60. Elsevier.

Nowak, K.; Webb, R.; Cifelli, R.; and Brekke, L. 2017. Sub-seasonal climate forecast rodeo. In *2017 AGU Fall Meeting, New Orleans, LA, 11-15 Dec.*

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.; Bharwani, S.; MacLeod, D.; James, R.; Fleming, L.; Morse, A. P.; Eggen, B.; Graham, R.; Kjellstrom, E.; Becker, E.; Pegion, K. V.; Holbrook, N. J.; McEvoy, D.; Depledge, M.; Perkins-Kirkpatrick, S.; Brown, T. J.; Street, R.; Jones, L.; Remenyi, T. A.; Hodgson-Johnston, I.; Buontempo, C.; Lamb, R.; Meinke, H.; Arheimer, B.; and Zebiak, S. E. 2017. Potential applications of subseasonal-to-seasonal (s2s) predictions. *Meteorological Applications* 24(3):315–325.