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

Paulo Orenstein*, Jessica Hwang*, Karl Pfeiffer†, Judah Cohen† and Lester Mackey‡

* Stanford University, † Atmospheric and Environmental Research, ‡ Microsoft Research

Summary

To improve the accuracy of long-term weather forecasts, the Bureau of Reclamation (USBR) 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 predict temperature and precipitation in the western U.S. two to four weeks and four to six weeks in advance. Our machine learning solution is an ensemble of two models: (i) `MultiLLR`, a local linear regression with multitask model selection; and (ii) `AutoKNN`, a multitask k-nearest neighbor autoregression. The ensemble significantly outperforms the government baselines as well as the top Rodeo competitor for each target variable and forecast horizon. We also release our `SubseasonalRodeo` dataset, collected to train and evaluate our forecasting system.

Forecast Rodeo

Goal Water managers in the western U.S. rely on subseasonal forecasts of temperature and precipitation to prepare for droughts and other wet weather extremes. This motivated the USBR and NOAA to conduct the Subseasonal Climate Forecast Rodeo: a real-time forecasting competition in which, every two weeks, contestants submitted four forecasts: for each variable

- **temp**: average temperature (°C)
- **prec**: total precipitation (mm),

predict it at each of two forecast horizons,

- **weeks 3-4**: 15-28 days ahead
- **weeks 5-6**: 29-42 days ahead.

The geographic region of interest is shown in Figure 1, with a total of $G = 514$ grid points. The contest ran from April 18, 2017 to April 3, 2018.

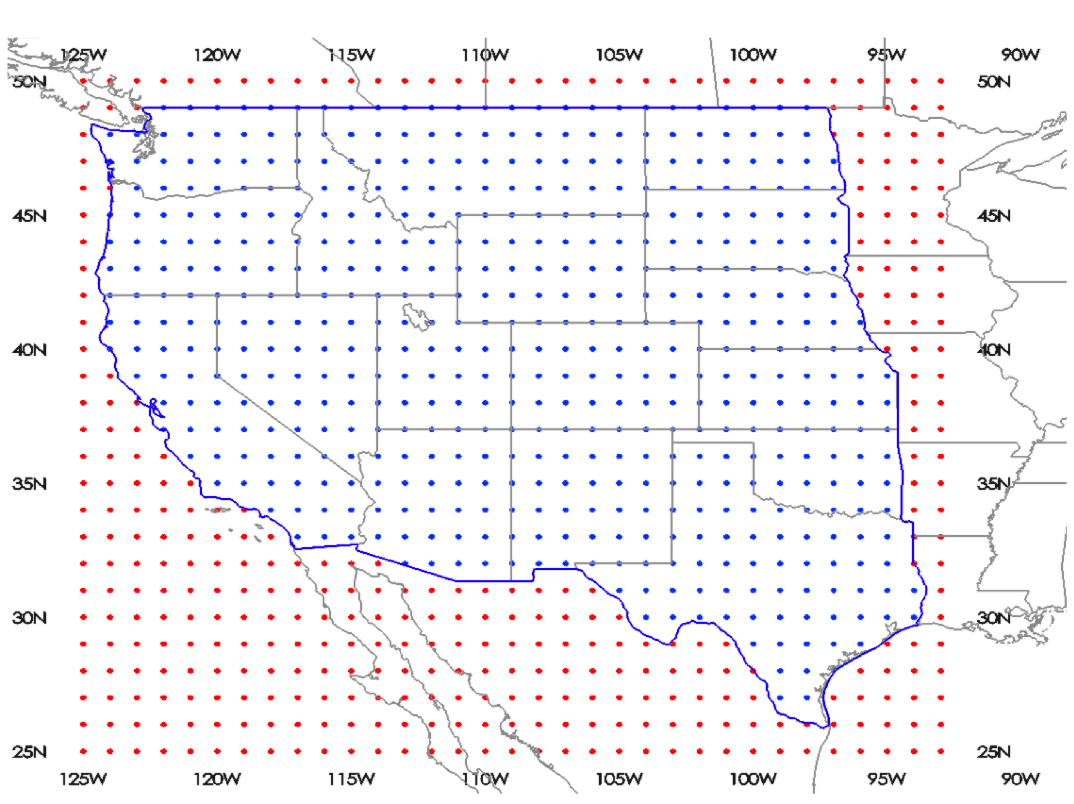

Fig 1: Forecast Rodeo prediction region.

Evaluation Let t be a date, and let $\texttt{year}(t)$, $\texttt{doy}(t)$, and $\texttt{monthday}(t)$ denote the associated year, day of the year, and month-day combination (e.g., January 1). For the two-week period beginning on t, associate an observed average temperature or total precipitation $\mathbf{y}_t \in \mathbf{R}^G$ and an observed *anomaly* and *climatology*

$$\mathbf{a}_t = \mathbf{y}_t - \mathbf{c}_{\texttt{monthday}(t)}, \qquad \mathbf{c}_d \triangleq \frac{1}{30} \sum_{\substack{t:\,\texttt{monthday}(t)=d,\\ 1981\le\texttt{year}(t)\le 2010}} \mathbf{y}_t.$$

Contestants were judged on the *skill* between their forecast anomalies $\hat{\mathbf{a}}_t = \hat{\mathbf{y}}_t - \mathbf{c}_{\texttt{monthday}(t)}$ and observed anomalies:

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

The overall contest skill was the mean over all dates.

Baselines To qualify for a prize, contestants had to achieve higher mean skill than two government benchmarks: (i) a debiased 32-member ensemble version of the operational, physics-based U.S. Climate Forecasting System (CFSv2), and (ii) a damped persistence forecast.

`SubseasonalRodeo` Dataset

As the Rodeo did not provide data for training predictive models, we constructed our own dataset from a diverse collection of data sources (see [1] for details):

- **temperature**: daily maximum and minimum temperature measurements at 2 meters, in °C; the official contest target temperature variable is $\texttt{tmp2m} \triangleq \frac{\texttt{tmax}+\texttt{tmin}}{2}$;
- **precipitation**: daily precipitation, in mm;
- **sea surface temperature (SST), sea ice concentration (SIC)**: daily; top three principal components for grid points in the Pacific basin region;
- **Multivariate ENSO index (MEI)**: bimonthly values; MEI is a scalar summary of the El Niño/Southern Oscillation, an ocean-atmosphere coupled climate mode;
- **Madden-Julian oscillation (MJO)**: daily values of phase and amplitude; MJO is a metric of tropical convection influencing western U.S. subseasonal climate;
- **Relative humidity and pressure**: daily relative humidity and pressure near the surface (sigma level 0.995);
- **Geopotential height**: daily; top three principal components for height at which 10mb of pressure occurs;
- **North American Multi-Model Ensemble (NMME)**: monthly mean forecast for each model in a collection of 10 physics-based models were averaged according to the number of target period days that fell in each month; the predictions from each model were then averaged.

Models

Our forecasting system relies on two regression models.

`MultiLLR`: multitask model selection with full data.

- *predictors*: lagged measurements from all data sources in the `SubseasonalRodeo` dataset;
- *method*: multitask backward stepwise procedure adapted to cosine objective, built atop per-grid-point local linear regression (with locality determined by day of the year).

`AutoKNN`: autoregression model with kNN features.

- *predictors*: lagged measurements of target variable and anomalies of the 20 most similar historical dates (similarity measured as skill between anomaly trajectory leading up to target date and up to each historical date);
- *method*: weighted local linear regression, with weights given by 1/(variance of the target anomaly vector).

Ensembling For a given target date, the ensemble forecast anomalies are the average of the ℓ_2-normalized predicted anomalies of the two models:

$$\hat{\mathbf{a}}_{\text{ensemble}} \triangleq \frac{1}{2} \frac{\hat{\mathbf{a}}_{\text{multillr}}}{\|\hat{\mathbf{a}}_{\text{multillr}}\|_2} + \frac{1}{2} \frac{\hat{\mathbf{a}}_{\text{autoknn}}}{\|\hat{\mathbf{a}}_{\text{autoknn}}\|_2}.$$

Prop. Skill of $\hat{\mathbf{a}}_{\text{ensemble}}$ is better than the average skill of $\hat{\mathbf{a}}_{\text{multillr}}$ and $\hat{\mathbf{a}}_{\text{autoknn}}$ whenever that average skill is positive.

Fig 2: Prec 3-4: Distribution of the month of the most similar neighbor learned by `AutoKNN` as a function of month of target date.

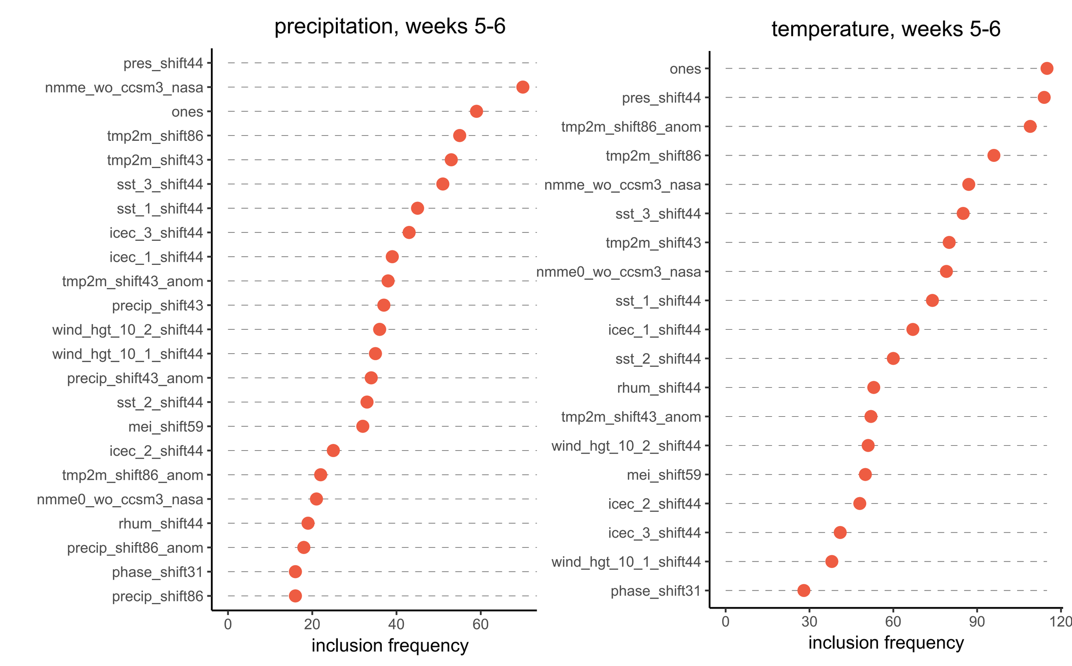

Fig 3: Frequency with which each candidate feature was selected in `MultiLLR`. Pressure, temperature, intercept and NMME are common.

Results

Contest-period skills The table below shows the average skills for each of our methods and each of the baselines over the contest period, April 2017 to April 2018. All three of our methods (`MultiLLR`, `AutoKNN`, ensemble) outperform the official contest baselines (debiased CFSv2 and damped persistence), and our ensemble outperforms the top Rodeo competitor in all four prediction tasks.

task	multillr	autoknn	ens	cfsv2	damp	top
temp 3-4	0.285	0.280	**0.341**	0.158	0.195	0.2855
temp 5-6	0.237	0.281	**0.307**	0.219	-0.076	0.235
prec 3-4	0.167	0.215	**0.238**	0.071	-0.146	0.214
prec 5-6	0.221	0.187	**0.241**	0.022	-0.161	0.216

Historical forecast evaluation In the period from April 2011 to April 2017, `AutoKNN` and the ensembled model surpass the reconstructed debiased CFSv2 baseline. Further ensembling with CFSv2 gives even better skill.

task	multillr	autoknn	ens	cfsv2	ens-cfsv2
temp 3-4	0.223	0.311	0.307	0.256	**0.351**
temp 5-6	0.220	0.281	0.296	0.214	**0.328**
precip 3-4	0.157	0.151	0.189	0.086	**0.196**
precip 5-6	0.131	0.140	0.170	0.069	**0.176**

Fig 4: Performance of ens, cfs, and ens-cfs in the historical period.

Conclusion Main takeaways from our work:

- Statistical models offer significant subseasonal forecast improvements over operational physics-based models.
- Ensembling statistical and physics-based forecasts can produce further improvements.
- New `SubseasonalRodeo` dataset for training and evaluation (`https://doi.org/10.7910/DVN/IHBANG`).

Main References

[1] J Hwang, P Orenstein, K Pfeiffer, J Cohen, and L Mackey. 2018. Improving Subseasonal Forecasting in the Western US with Machine Learning. *arXiv:1809.07394*.

[2] K Nowak, RS Webb, R Cifelli, and LD Brekke. 2017. Sub-Seasonal Climate Forecast Rodeo. In *2017 AGU Fall Meeting, New Orleans, LA, 11-15 Dec.*