Learned Benchmarks for Subseasonal Forecasting

Soukayna Mouatadid[1], Paulo Orenstein[2], Genevieve Flaspohler[3,4], Miruna Oprescu[5], Judah Cohen[6,7], Franklyn Wang[8], Sean Knight[9], Maria Geogdzhayeva[10], Sam Levang[11], Ernest Fraenkel[12], Lester Mackey[5]

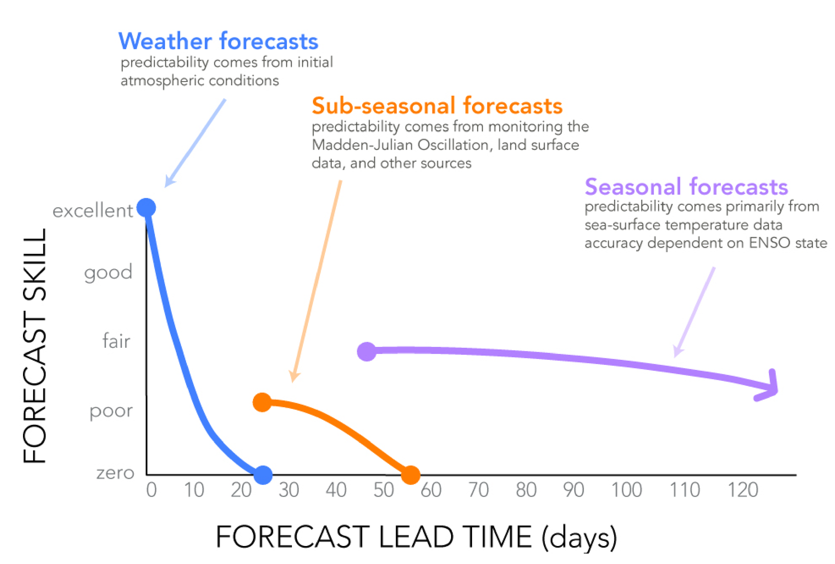
[1] Department of Computer Science, University of Toronto; [2] Instituto de Matemática Pura e Aplicada; [3] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology; [4] Department of Applied Ocean Science and Engineering, Woods Hole Oceanographic Institution; [5] Microsoft Research New England; [6] Atmospheric and Environmental Research; [7] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology; [8] Department of Mathematics, Harvard University; [9] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology; [10] Department of Physics, Massachusetts Institute of Technology; [11] Salient Predictions Inc.; [12] Department of Biological Engineering, Massachusetts Institute of Technology

INTRODUCTION

Subseasonal weather prediction (3-6 weeks ahead) is a crucial pre-requisite for:

- Preparing droughts and floods
- Agriculture planning
- Allocation of water resources
- Managing wildfires

It is a challenging forecast horizon for both meteorological and ML models:



Source: https://iri.columbia.edu/news/qa-subseasonal-prediction-project/

Objective:

- We develop a toolkit of subseasonal models that outperform operational weather models as well as state-of-the-art learning methods from the literature.

FORECASTING TASKS

- **Target variables:** Average temperature and Accumulated precipitation
- **Lead times:** Weeks 3-4 ahead and Weeks 5-6 ahead
- **Geographical region:** U.S., 1°x1° resolution, G = 862 gridpoints
- **Loss function:** RMSE, skill
- **Dataset:** Improved SubseasonalRodeo dataset (Hwang et al., 2019)

MODELS

Baselines:

- **Climatology:** average temperature or precipitation for specific day and month over 1981-2010.
- **CFSv2:** operational U.S. physics-based model from NCEP.
- **Persistence:** predict most recent value.

Learning models:

- **AutoKNN,** introduced in (Hwang et al., 2019)
- **Informer,** introduced in (Zhou, 2021)
- **LocalBoosting,** introduced in (Prokhorenkova et al., 2018)
- **MultiLLR,** introduced in (Hwang et al., 2019)
- **N-BEATS,** introduced in (Orenshkin, 2020)
- **Prophet,** introduced in (Taylor and Letham, 2018)
- **Salient 2.0,** introduced in (Schmitt, 2019)

Our toolkit:

- **Climatology++:** Use adaptively selected window around target day for averaging.
- **CFSv2++:** Average over range of issuance date and lead times, adaptively debiasing using selected window.
- **Persistence++:** Learned combination of lagged measurements with NWP.

ENSEMBLING

Uniform ensemble:

- Average over base models
- Typical solution in the weather community

Online ensemble:

- Runs a follow-the-regularized-leader online learning method
- Results in an adaptive convex combination of base models

Base models:

- Climatology++, CFSv2++, Persistence++

RESULTS

Table 1: Average percentage skill and percentage improvement over mean debiased CFSv2 RMSE across 2011-2020 in the contiguous U.S. The best performing model in each model group is bolded, and the best performing model overall is shown in green.

		% Improvement over Mean Deb. CFSv2 RMSE				Average % Skill			
		Temperature		Precipitation		Temperature		Precipitation	
Group	Model	weeks 3-4	weeks 5-6	weeks 3-4	weeks 5-6	weeks 3-4	weeks 5-6	weeks 3-4	weeks 5-6
Baselines	Climatology	**0.13**	**2.93**	**7.79**	**7.51**	–	–	–	–
	Debiased CFSv2	–	–	–	–	**24.94**	**19.12**	5.77	4.28
	Persistence	−109.94	−170.1	−28.27	−31.92	10.64	6.22	**8.31**	**7.41**
Toolkit	Climatology++	2.06	4.83	**8.86**	**8.57**	18.61	18.87	15.04	14.99
	CFSv2++	5.94	**7.09**	8.37	8.06	32.38	**29.19**	**16.34**	**16.09**
	Persistence++	**6.00**	6.43	8.61	7.89	**32.4**	26.73	13.38	9.77
Learning	AutoKNN	0.93	3.22	7.73	7.33	12.43	8.56	6.66	5.93
	Informer	−40.61	−39.57	−2.05	−2.53	0.55	0.01	6.15	5.86
	LocalBoosting	−0.76	−0.29	7.36	6.89	14.44	12.69	10.82	9.72
	MultiLLR	**2.45**	2.21	7.12	6.65	**24.5**	16.68	9.49	7.97
	N-BEATS	−46.71	−52.05	−19.19	−21.32	9.21	4.16	5.48	4.46
	Prophet	1.13	**3.78**	**8.42**	**8.12**	20.21	**19.78**	**13.51**	**13.41**
	Salient 2.0	−6.95	−4.05	2.97	2.65	11.24	11.77	10.11	9.99
Ensembles	Uniform Toolkit	6.47	7.55	9.47	9.05	33.58	30.56	18.94	18.35
	Online Toolkit	6.67	7.67	9.51	9.04	33.27	30.06	18.86	17.91

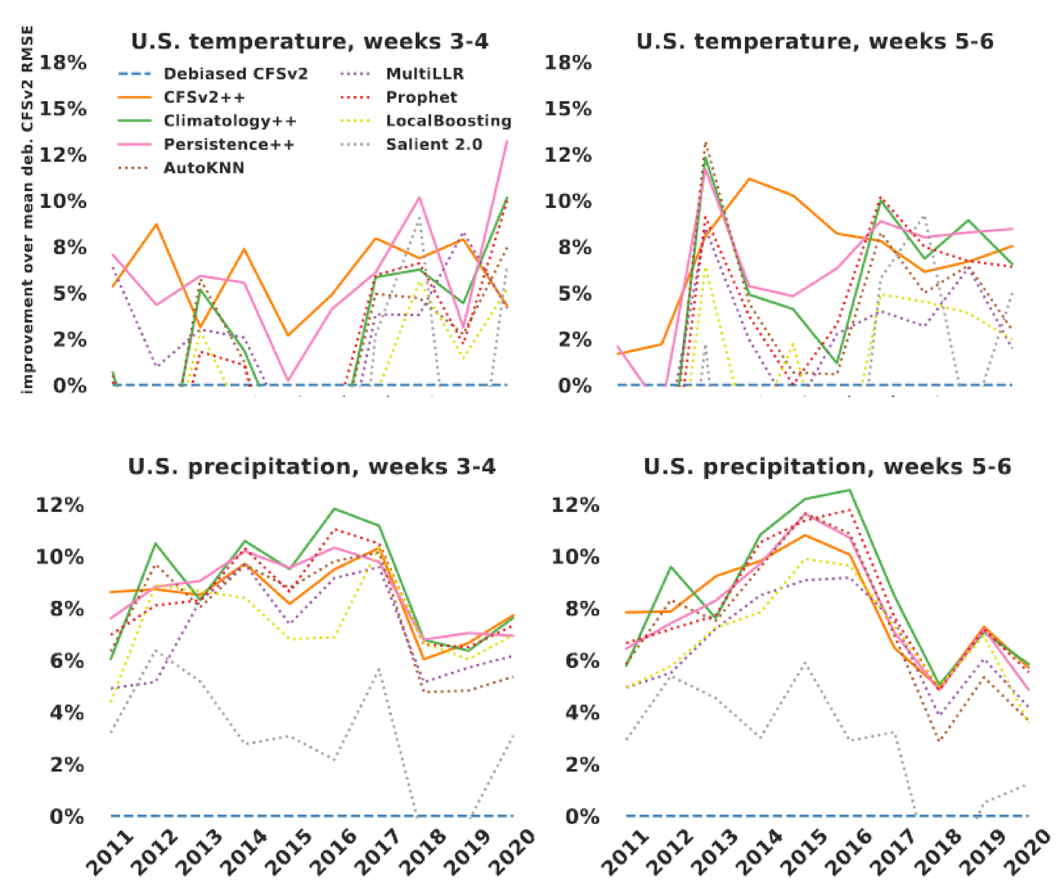


Figure 1: Per season and per year improvement over mean debiased CFSv2 RMSE across the contiguous U.S. and the years 2011-2020. Despite their simplicity, the toolkit models (solid lines) consistently outperform debiased CFSv2 and the state-of-the-art learners (dotted lines).

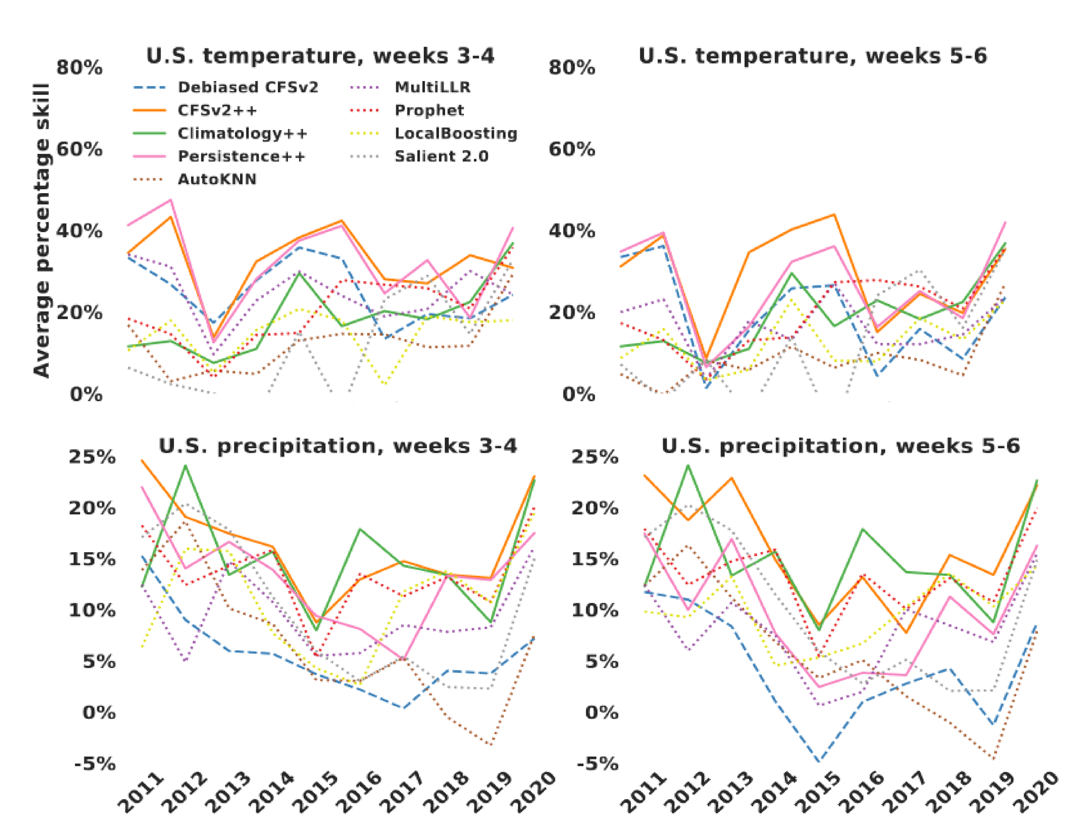


Figure 2: Per season and per year average skill across the contiguous U.S. and the years 2011-2020. Despite their simplicity, the toolkit models (solid lines) consistently outperform debiased CFSv2 and the state-of-the-art learners (dotted lines).

COMPARING TO ECMWF

Table 2: Average percentage skill and percentage improvement over mean debiased CFSv2 RMSE across 2016-2020 in the contiguous U.S. The best performing model in each model group is bolded, and the best performing model overall is shown in green.

		% Improvement over Mean Deb. CFSv2 RMSE				Average % Skill			
		Temperature		Precipitation		Temperature		Precipitation	
Group	Model	weeks 3-4	weeks 5-6	weeks 3-4	weeks 5-6	weeks 3-4	weeks 5-6	weeks 3-4	weeks 5-6
Baselines	Climatology	**1.56**	**3.92**	**8.7**	**7.56**	–	–	–	–
	Debiased CFSv2	–	–	–	–	**22.64**	**15.71**	2.84	1.68
	Persistence	−105.57	−169.22	−28.05	−33.43	9.12	2.27	**8.11**	**6.21**
Toolkit	Climatology++	3.88	6.44	**9.79**	**8.61**	22.09	23.2	**15.34**	**15.06**
	CFSv2++	5.65	6.65	8.94	7.6	30.91	26.87	14.6	13.85
	Persistence++	**7.06**	**7.86**	9.06	7.57	**31.46**	**28.04**	10.03	6.61
ECMWF	Debiased Control	−29.05	−33.25	−30.81	−31.84	18.52	13.71	0.82	3.17
	Debiased Ensemble	**4.62**	**3.69**	**7.90**	**6.41**	**32.27**	**26.61**	**13.12**	**9.10**
Ensembles	Uniform Toolkit	7.43	8.27	10.04	8.77	32.77	29.75	16.53	15.71
	Online Toolkit	7.2	7.96	10.08	8.62	32.22	28.38	17.19	15.42

SPATIAL IMPROVEMENT AND BIAS

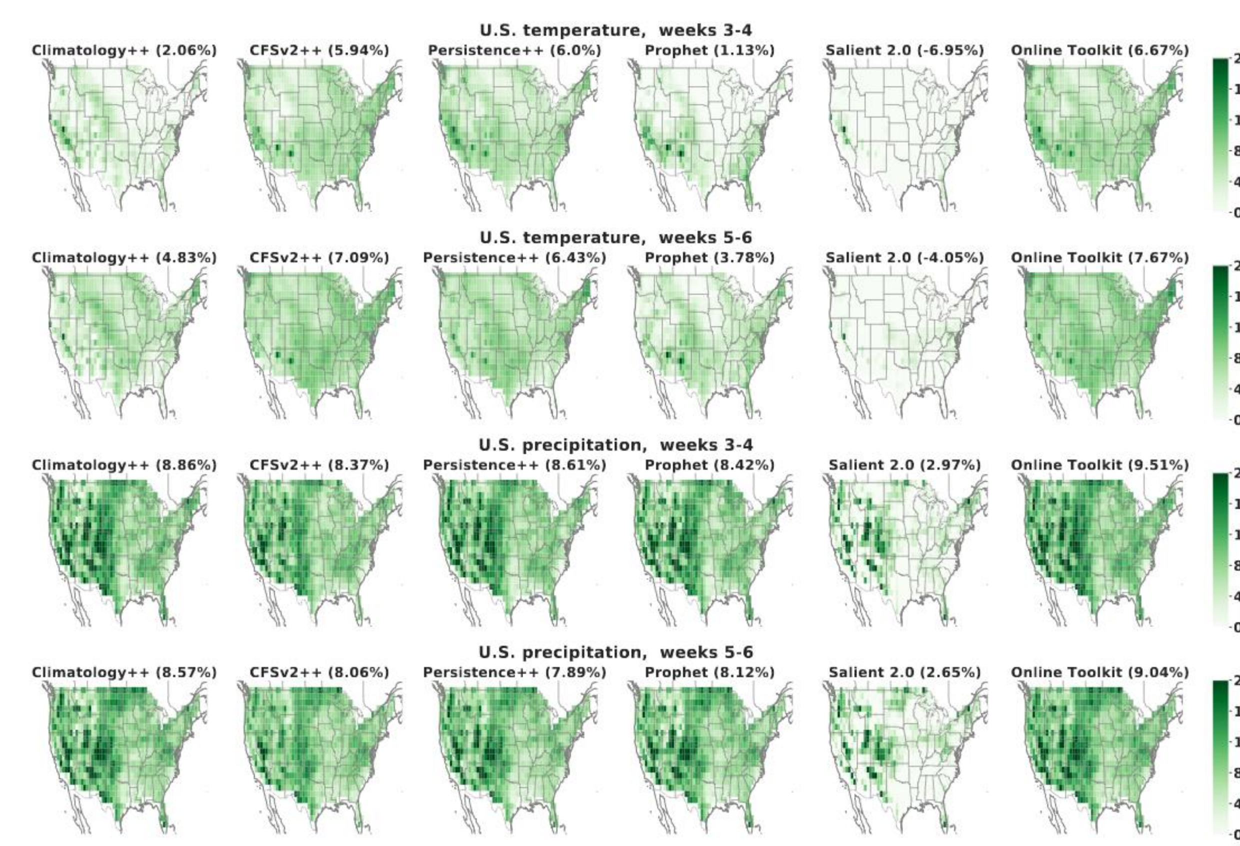


Figure 3: Percentage improvement over mean debiased CFSv2 RMSE in the contiguous U.S. over 2011-2020. White grid points indicate negative or 0% improvement.

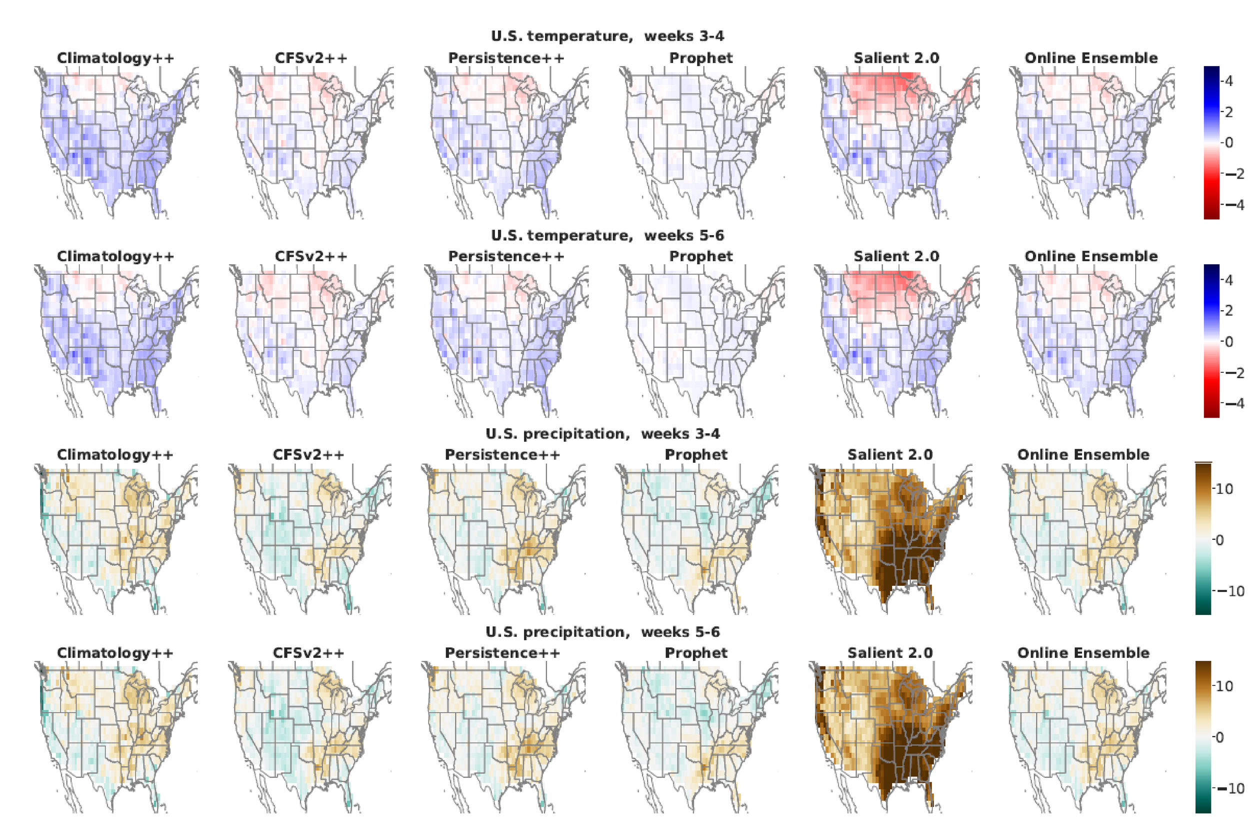


Figure 4: Model bias in the contiguous U.S. over 2011-2020. White grid points indicate zero bias.

WESTERN U.S. COMPETITION

Table 3: Percentage improvement over mean debiased CFSv2 RMSE over 26 contest dates (2019-2020) in the Western U.S. The best performing models within each class of models are shown in bold, while the best performing models overall are shown in green.

Group	Model	Temp. weeks 3-4	Temp. weeks 5-6	Precip. weeks 3-4	Precip. weeks 5-6
Contest baselines	Salient	–	–	11.10	7.02
	Climatology	10.22	−0.76	5.82	2.25
Contestants	1^{st} place	17.12	**8.47**	**11.54**	8.63
	2^{nd} place	16.67	7.04	11.10	8.03
	3^{rd} place	15.47	6.90	10.62	7.94
Learning	AutoKNN	13.09	2.90	7.50	3.05
	LocalBoosting	12.85	4.09	7.25	3.71
	MultiLLR	9.54	1.12	8.95	4.58
	Prophet	**15.68**	**6.86**	6.88	3.40
	Salient 2.0	11.15	2.91	12.65	**8.56**
Toolkit	Climatology++	15.54	6.43	8.35	4.69
	CFSv2++	6.67	**9.26**	**8.70**	**5.51**
	Persistence++	**16.59**	8.27	8.20	4.51
Ensembles	Uniform Toolkit	14.96	9.58	9.31	5.89
	Uniform Toolkit + Learning	15.89	8.79	10.43	6.79
	Online Toolkit	**16.71**	8.70	8.85	5.19
	Online Toolkit + Learning	14.70	7.97	**12.52**	**8.18**

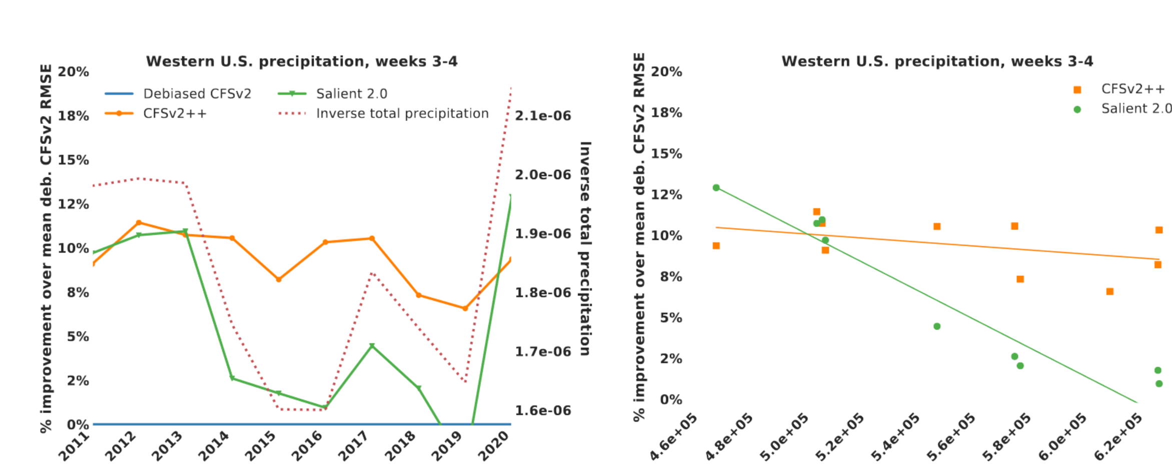


Figure 5: Temporal plot (left) and scatter plot (right) of yearly total precipitation and percentage improvement over mean debiased CFSv2 RMSE in the Western U.S. across 2011-2020

References

- Hwang, J., Orenstein, P., Cohen, J., Pfeiffer, K., & 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 (pp. 2325–2335).
- Oreshkin, B. N., Carpov, D., Chapados, N., & Bengio, Y. (2020). N-BEATS: neural basis expansion analysis for interpretable time series forecasting. In 8th international conference on learning representations, ICLR 2020, addis ababa, ethiopia, april 26-30, 2020. OpenReview.net. Retrieved from https://openreview.net/forum?id=r1ecqn4YwB
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). Catboost: unbiased boosting with categorical features. In Advances in neural information processing systems (pp. 6638–6648).
- Schmitt, R. (2019). Salient predictions: Validation summary. https://storage.googleapis.com/content.salientpredictions.com/Salient%20Validation%20Summary.pdf. (Accessed: 2021-05-29)
- Taylor, S. J., & Letham, B. (2018). Forecasting at scale. The American Statistician, 72(1), 37–45.
- Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. In The thirty-fifth AAAI conference on artificial intelligence, AAAI 2021 (p. online). AAAI Press.