An Accurate and Scalable Subseasonal Forecasting Toolkit for the United States

Soukayna Mouatadid [1] **Paulo Orenstein** [2] **Genevieve Flaspohler** [3] **Miruna Oprescu** [4] **Judah Cohen** [5]
Franklyn Wang [6] **Sean Knight** [3] **Ernest Fraenkel** [3] **Lester Mackey** [4]

Abstract

We develop a subseasonal forecasting toolkit of accurate and highly scalable benchmarks that outperform both the United States operational Climate Forecasting System (CFSv2) and state-of-the-art learning methods from the literature. Our new learned benchmarks include (a) Climatology++, an enhanced form of climatology using knowledge of only the day of the year; (b) CFSv2++, a learned correction for CFSv2; and (c) Persistence++, an augmented persistence model that combines lagged measurements with CFSv2 forecasts. These methods alone improve upon CFSv2 accuracy by 9% for US precipitation and 6% for US temperature over 2011-2020. Ensembling our benchmarks with diverse forecasting methods leads to even further gains. Overall, we find that augmenting classical forecasting approaches with learned corrections yields an effective, low-cost strategy for building next-generation subseasonal forecasting models.

1. Introduction

It is estimated that 2.7 trillion dollars of the US economy is sensitive to the impacts of weather and climate [1]. Improving our forecasting skill is of interest to all sectors of the economy. Recently, seasonal climate forecasts have become an important aspect of policy and decision-making and are utilized in a broad range of applications [2].

While purely physics-based numerical weather prediction dominates the landscape of short-term weather forecasting [3–6], such deterministic methods have a limited skillful (i.e., accurate) forecast horizon due to the chaotic nature of weather [7]. In addition, the utility of longer-lead subseasonal and seasonal outlooks, which depend on both local weather and global climate influences, is still limited by skill in dynamical forecasting methods [8].

Learning techniques from machine learning and statistics have been leveraged in subseasonal forecasting [9–18]. While progress has been made in applying black-box techniques to traditional meteorological models, we show that applying select learning techniques leads to a scalable toolkit of models that can outperform current operational weather benchmarks as well as state-of-the-art learning models. This suggests building on classical models is a simple and scalable strategy for better subseasonal forecasting.

2. Forecasting Tasks

To evaluate the performance of subseasonal climate models, we consider forecasts of two variables: average temperature (°C) and accumulated precipitation (mm) over a two-week period. These variables are forecasted at two time horizons: 15-28 days ahead (weeks 3-4) and 29-42 days ahead (weeks 5-6). The geographical region for these predictions is the contiguous US, delimited by latitudes 25N to 50N and longitudes 125W to 67W, at a 1° by 1° resolution. The contiguous US has a total of $G = 862$ grid points.

These specific time frames and geographical region are used by the USBR and NOAA for their subseasonal decision making [11]. These forecasting tasks were motivated by the Subseasonal Forecast Rodeos I and II [19], year-long real-time competitions sponsored by USBR and NOAA to advance the state of subseasonal weather forecasting.

In the experiments below, models are trained with data up to 3-4 or 5-6 weeks before the 2-week target period. The test set is taken as the years 2011-2020. Hyperparameters are picked using the three years previous to the target period. We evaluate both decade and year-long performance.

Forecasts are evaluated using root mean squared error (RMSE). For each date d, the RMSE is defined as

$$\text{rmse}_d = \sqrt{\frac{1}{G}\sum_{g=1}^{G}(\hat{y}_{d,g} - y_{d,g})^2}. \tag{1}$$

where $y_{d,g}$ (respectively, $\hat{y}_{d,g}$) denotes the ground-truth measurement (respectively, predicted value) for grid point g and the two-week period starting on date d. Over a given set of dates, the error is given by the average RMSE.

3. Dataset

The features available to our models and baselines are collected in the SUBSEASONALCLIMATEUSA dataset. The ge-

ographical region is the continental USA at a $1^\circ \times 1^\circ$ spatial resolution, encompassing years 1980-2020. The variables included are: temperature, precipitation, sea surface temperature, sea ice concentration, multivariate ENSO index, Madden-Julian oscillation, relative humidity, geopotential heights and CFSv2. Variables are averaged over two-week periods. This dataset is an improved and upgraded version of the dataset introduced in [11].

4. Methods

4.1. Baselines

For all four tasks considered, we use three baseline models.

Climatology is the standard baseline for subseasonal forecasting, given by the average of the variable of interest for a specific grid point, day and month over the years 1981-2010.

CFSv2 (Climate Forecasting System version 2) is one of the main operational physics-based model capable of subseasonal weather prediction, run by the National Centers for Environmental Prediction (NCEP) [20]. It is the main numerical weather prediction model baseline in the paper.

Persistence is a traditional forecasting model that predicts the most recently observed two-week target value [18], so the current weather information "persists" into the future.

4.2. Toolkit models

We enhance each of the baseline models by leveraging simple and effective statistical and machine learning techniques.

Climatology++ uses knowledge only of the day of the year: for a given target date, it predicts the vector of temperatures (respectively, precipitations) that minimizes mean RMSE (respectively, MSE) over the past 26 years and all days in an adaptively selected window around the target day of year.

CFSv2++ averages CFSv2 forecasts over a range of issuance dates and lead times and then debiases the ensemble prediction for each grid cell and target date using an adaptively selected window around the target day of year.

Persistence++ fits least squares regression per grid point to combine lagged measurement forecasts (of 29 and 58 days for 3-4 week tasks, and 43 and 86 for 5-6 week tasks) with a CFSv2 ensemble forecast, optimally combining numerical weather forecasts with recent weather trends.

4.3. Learning models

We also consider seven state-of-the-art learning models.

AutoKNN, as described in [11], is a weighted local linear regression with features derived from historical measurements of the target variable (temperature or precipitation).

MultiLLR, proposed in [11], is a local linear regression model with multitask feature selection. For a target date and grid point, it subsets the data to within 8 weeks of the target date and fits a backward stepwise linear regression to pick the best features, based on RMSE, over dates with the same month-day combination.

Prophet is an additive regression model for univariate time-series forecasting [21]. It captures seasonality by incorporating weekly and yearly seasonal trends on top of a piecewise linear or logistic growth curve. This model is one of the winning solutions in the Subseasonal Forecast Rodeo II [19].

LocalBoosting is a boosting model based on CatBoost [22]. For each grid point, it uses as features the values of all the weather variables in the SUBSEASONALCLIMATEUSA dataset on a geographic region around the grid point, thus allowing it to use neighboring spatial information.

N-BEATS [23] is a neural network based model, inspired by the ResNet [24]. It successively residualizes the data against current predictions, allowing it to capture ever more complicated patterns. N-BEATS is trained on the univariate time series at each grid point.

Informer [25] is a transformer-based deep model [26] which has attained great performance on short-range weather forecasting. It processes very long patterns, and uses sparse connections to overcome computational bottlenecks. Informer is a multivariate time-series model, so a single model is trained for all grid points.

Salient 2.0 is based on Salient [27], the winning solution for the Subseasonal Forecast Rodeo I [19]. It consists of an ensemble of feed-forward fully-connected neural networks, trained on sea surface temperature (SST) data. Salient 2.0 uses 50 randomly-initialized networks trained on SST, multivariate ENSO and MJO data and selects 10 networks based on a new data split strategy described in Section 2.

4.4. Ensembling

Ensembling is a powerful technique for subseasonal weather forecasting [11; 28]. We consider two ways of ensembling: Uniform Emsemble and Online Ensemble, each of which uses a set of six base models: Climatology++, CFSv2++, Persistence++, LocalBoosting, MultiLLR and Salient 2.0.

Uniform Ensemble, inspired by methods in use in the weather community, simply takes uniform averages over the predictions of a set of models.

Online Ensemble runs a follow-the regularized-leader online learning algorithm over the base models for the validation period to produce a weight for each model, and finally outputs the weighted average of predictions [29]. The result is an adaptive convex combination of different base models.

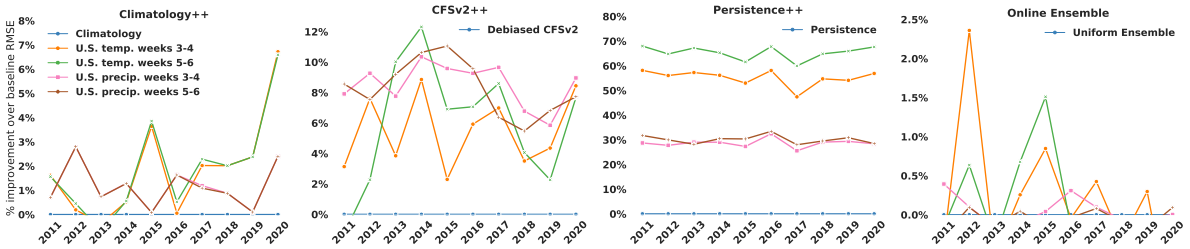

Figure 1. Relative percentage improvement of toolkit models over their traditional counterpart in terms of RMSE, for temperature and precipitation, 3-4 and 5-6 weeks ahead, over 2011-2020. Toolkit models are consistently better than their classical versions.

Table 1. Percentage improvement over debiased CFSv2 RMSE for 2011-2020 in the US.

	Toolkit			Learning							Ensembles	
	Clim++	CFSv2++	Pers++	AKNN	LBoost	Inform.	MLLR	N-Beats	Prophet	Sal. 2.0	Uniform	Online
Temp. 3-4w	1.60	5.49	5.60	0.51	−1.18	−40.58	2.04	−47.33	0.71	−9.28	6.07	**6.23**
Temp. 5-6w	3.90	6.16	5.51	2.26	−1.28	−65.27	1.24	−53.55	2.83	−4.98	6.64	**6.79**
Precip. 3-4w	9.03	8.53	8.78	7.90	7.53	0.83	7.29	−18.97	8.59	3.17	9.63	**9.69**
Precip. 5-6w	8.85	8.34	8.17	7.62	7.17	0.49	6.94	−20.95	8.40	2.96	**9.33**	9.27

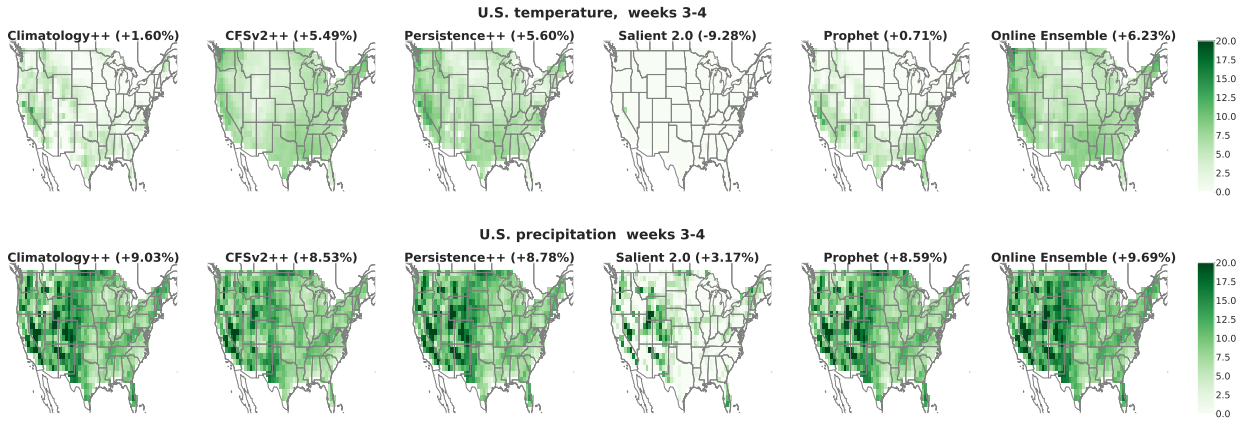

Figure 2. Percentage improvement over debiased CFSv2 RMSE for temperature and precipitation, weeks 3-4 in the US over 2011-2020. Models display different levels of improvement over different geographical regions in the contiguous US.

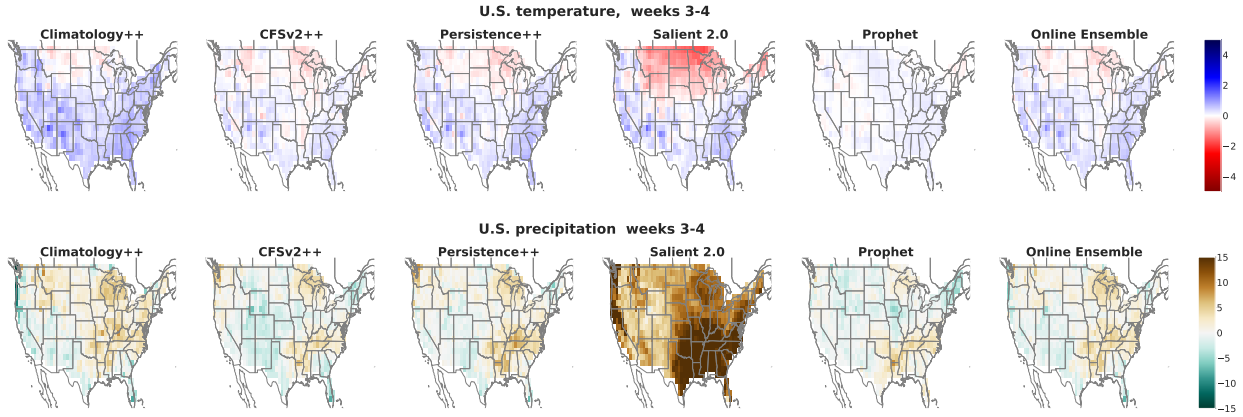

Figure 3. Average model bias for temperature and precipitation, weeks 3-4 in the US, over 2011-2020. From the top row, most models show a cold bias in the south, while Salient exhibits a strong warm bias in the north. Models show wet biases, except for Salient.

Table 2. Percentage improvement over debiased CFSv2 RMSE for 2011-2020 in the US for classical models vs toolkit. The CFSv2 column refers to the raw CFSv2 prior to debiasing.

	CLIM	CLIM++	CFSV2	CFSV2++	PERS	PERS++
TEMP. 3-4W	-0.29	**1.60**	-14.17	**5.49**	-110.83	**5.60**
TEMP. 5-6W	1.97	**3.90**	-15.37	**6.16**	-172.76	**5.51**
PRECIP. 3-4W	7.96	**9.03**	-4.57	**8.53**	-28.03	**8.78**
PRECIP. 5-6W	7.79	**8.85**	-4.83	**8.34**	-31.51	**8.17**

5. Results

In this section, we evaluate our individual models as well as the ensemble models over the the period 2011-2020. We find that the enhancements made to the toolkit models lead to better performance relative to the traditional weather models they are based on, as well as state-of-the-art learning alternatives. Also, most models show heterogeneous levels of accuracy in both time and space, motivating the idea of developing a suite of models. Finally, we show that properly ensembling can yield substantial gains.

To highlight the improvement of individual toolkit models over their traditional counterparts, Figure 1 shows the yearly relative percentage improvement of the toolkit models (Climatology++, CFSv2++, and Persistence++) relative to the baselines (Climatology, CFSv2, and Persistence) for each model [1]. Climatology++, CFSv2++, and Persistence++ consistently outperform Climatology, CFSv2, and Persistence, respectively, on all four tasks considered. Table 2 gives a more detailed comparison over 2011-2020: while the baselines were oftentimes worse than CFSv2 in terms of RMSE, all the toolkit models managed to overcome these shortcomings to give consistently better performances. The result is particularly noticeable for Persistence++, showing that the addition of numerical weather forecasting and well-chosen target lags can significantly improve prediction.

Furthermore, Table 1 shows that the toolkit models generally outperform modern learning models. In fact, the only ones able to overcome the worst of the toolkit models were Multi-LLR (for temperature) and Prophet (for precipitation), both of which involve few degrees of freedom. Thus, it seems that the flexibility of boosting and deep learning methods seem of limited use in the subseasonal setting.

Given the different approaches taken by the toolkit models, one may wonder if they are each able to capture different but complementary signals. One way to test this is to look at the individual data points for which each algorithm performs well. Figure 1 shows that the performance of models often varies over time, and Figure 2, which displays the percentage improvement in RMSE per grid cell, shows that performance also varies over space. Generally, the data supports the notion that these models are capturing slightly different signals.

Further, another instance of performance heterogeneity can be seen in Figure 3, which shows spatial patterns of average model bias. The average bias maps for temperature show a cold bias over the southern half of the US for Climatology++, Persistence++, Salient 2.0 and Online Ensemble, whereas Salient 2.0 shows an additional warm bias in the center north. For precipitation, all models, save for Salient 2.0, show wet biases in the western half of the US and dry biases in the eastern half, while Salient 2.0 displays a strong dry bias over the US that is pronounced in the eastern half.

Thus, considering how models are able to capture different signals, we turn to the question of combining such signals. Table 1 shows that, indeed, our highest-performing models are both ensembles. Note that even the simple averaging strategy of Uniform Ensemble is enough to drastically increase the performance of the final model. On the other hand, the more thoughtful hinting strategy employed by Online Ensemble is able to better combine the individual model predictions. In particular, Figure 1 shows that the performance of Online Ensemble is generally stronger than that of the Uniform Ensemble on all tasks. This remains useful even despite the glaring shortcomings of Salient 2.0 in the precipitation forecasting: adding it to the online ensemble actually improved performance for both 3-4 week and 5-6 week lead times.

6. Conclusion

In this work, we developed a toolkit of accurate and scalable benchmark models for sub-seasonal forecasting of temperature and precipitation in the US by applying simple statistical tools to classical subseasonal weather forecasting models. The toolkit models displayed better performance versus their classical counterpart as well as state-of-the-art learning alternatives. We also showed that prediction accuracy can vary significantly in time and space, and that no single model seems to dominate the subseasonal landscape. With this in mind, we showed that ensembling heterogeneous models was one way to improve predictions; in particular, by considering ensembling as an online learning problem lead to significant gains.

Overall, we found that these simple strategies for combining physics and data-driven models lead to powerful yet scalable subseasonal forecasting models. We anticipate these insights and improvements will benefit both researchers and practicioners in benchmarking as well as creating next-generation subseasonal forecasting models.

[1] The percentage improvement axis in Figure 1 was clipped at zero, in order to highlight the positive percentage change RMSE over debiased CFSv2. For all four models presented, the minimum percentage improvement was less than or equal to -1.5%.

References

[1] NOAA, "Noaa economic statistics. office of policy and strategic planning," *US Department of Commerce*, vol. 26, 2002.

[2] A. Troccoli, "Seasonal climate forecasting," *Meteorological Applications*, vol. 17, p. 251–268, 2010.

[3] A. G. Barnston, M. K. Tippett, M. L. L'Heureux, S. Li, and D. G. DeWitt, "Skill of real-time seasonal enso model predictions during 2002–11: is our capability increasing?," *Bulletin of the American Meteorological Society*, vol. 93, p. 631–651, 2012.

[4] F. Doblas-Reyes, J. Garcia-Serrano, F. Lienert, F. Biescas, and L. Rodrigues, "Seasonal climate predictability and forecasting: status and prospects," *WIREs Climate Change*, vol. 4, p. 245–268, 2013.

[5] a. M. National Academies of Sciences, Engineering, *Next generation earth system prediction: strategies for subseasonal to seasonal forecasts*. The National Academies Press, 2016.

[6] R. Reeves and D. Gemmill, *Climate Prediction Center: Reflections on 25 Years of Analysis, Diagnosis, and Prediction*. Washington, DC: US Government Printing Office, 2004.

[7] E. Lorenz, "Deterministic nonperiodic flow," *Journal of the Atmospheric Sciences*, vol. 20, no. 2, p. 130–141, 1963.

[8] A. Robertson, A. Kumar, M. Pea, and F. Vitart, "Improving and promoting subseasonal to seasonal prediction," *Bulletin of the American Meteorological Society*, vol. 96, p. 49– 53, 2015.

[9] J. Cohen, D. Coumou, J. Hwang, L. Mackey, P. Orenstein, S. Totz, and E. Tziperman, "S2s reboot: An argument for greater inclusion of machine learning in subseasonal to seasonal (s2s) forecasts," *WIREs Climate Change*, vol. 10, 2018.

[10] Y.-G. Ham, J.-H. Kim, and J.-J. Luo, "Deep learning for multi-year enso forecasts," *Nature*, vol. 573, no. 7775, pp. 568–572, 2019.

[11] J. Hwang, P. Orenstein, J. Cohen, K. Pfeiffer, and L. 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*, pp. 2325–2335, 2019.

[12] T. Arcomano, I. Szunyogh, J. Pathak, A. Wikner, B. R. Hunt, and E. Ott, "A machine learning-based global atmospheric forecast model," *Geophysical Research Letters*, vol. 47, no. 9, p. e2020GL087776, 2020.

[13] S. He, X. Li, T. DelSole, P. Ravikumar, and A. Banerjee, "Sub-seasonal climate forecasting via machine learning: Challenges, analysis, and advances," *arXiv preprint arXiv:2006.07972*, 2020.

[14] Q. F. Qian, X. J. Jia, and H. Lin, "Machine learning models for the seasonal forecast of winter surface air temperature in north america," *Earth and Space Science*, vol. 7, no. 8, p. e2020EA001140, 2020.

[15] S. Rasp, P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, and N. Thuerey, "Weatherbench: A benchmark data set for data-driven weather forecasting," *Journal of Advances in Modeling Earth Systems*, vol. 12, no. 11, p. e2020MS002203, 2020.

[16] A. Yamagami and M. Matsueda, "Subseasonal forecast skill for weekly mean atmospheric variability over the northern hemisphere in winter and its relationship to midlatitude teleconnections," *Geophysical Research Letters*, vol. 47, no. 17, p. e2020GL088508, 2020.

[17] D. Watson-Parris, "Machine learning for weather and climate are worlds apart," *Philosophical Transactions of the Royal Society A*, vol. 379, no. 2194, p. 20200098, 2021.

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

[19] K. Nowak, I. M. Ferguson, J. Beardsley, and L. D. Brekke, "Enhancing western united states subseasonal forecasts: Forecast rodeo prize competition series," in *AGU Fall Meeting 2020*, AGU, 2020.

[20] S. Saha, S. Moorthi, X. Wu, J. Wang, S. Nadiga, P. Tripp, D. Behringer, Y.-T. Hou, H.-y. Chuang, M. Iredell, *et al.*, "The ncep climate forecast system version 2," *Journal of climate*, vol. 27, no. 6, pp. 2185–2208, 2014.

[21] S. J. Taylor and B. Letham, "Forecasting at scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018.

[22] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," in *Advances in neural information processing systems*, pp. 6638–6648, 2018.

[23] B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio, "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, 2020.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[25] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "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, 2021.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA* (I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, eds.), pp. 5998–6008, 2017.

[27] R. Schmitt, "Salient predictions: Validation summary." https://storage.googleapis.com/content.salientpredictions.com/Salient%20Validation%20Summary.pdf, 2019. Accessed: 2021-05-29.

[28] J. Du, J. Berner, R. Buizza, M. Charron, P. L. Houtekamer, D. Hou, I. Jankov, M. Mu, X. Wang, M. Wei, *et al.*, "Ensemble methods for meteorological predictions," *Office note (National Centers for Environmental Prediction (U.S.))*, 2018.

[29] G. Flaspohler, F. Orabona, J. Cohen, S. Mouatadid, M. Oprescu, P. Orenstein, and L. Mackey, "Online learning with optimism and delay," in *International Conference on Machine Learning*, PMLR, 2021.