Adaptive Bias Correction for Improved Subseasonal Forecasting

Soukayna Mouatadid
University of Toronto
soukayna@cs.toronto.edu

Paulo Orenstein
Instituto de Matemática Pura e Aplicada

Genevieve Flaspohler
*n*Line Inc.
Massachusetts Institute of Technology
Woods Hole Oceanographic Institution

Judah Cohen
Atmospheric and Environmental Research
Massachusetts Institute of Technology

Miruna Oprescu
Cornell University

Ernest Fraenkel
Massachusetts Institute of Technology

Lester Mackey
Microsoft Research New England

Abstract

Subseasonal forecasting—predicting temperature and precipitation 2 to 6 weeks ahead—is critical for effective water allocation, wildfire management, and drought and flood mitigation. Recent international research efforts have advanced the subseasonal capabilities of operational dynamical models, yet temperature and precipitation prediction skills remains poor, partly due to stubborn errors in representing atmospheric dynamics and physics inside dynamical models. To counter these errors, we introduce an *adaptive bias correction* (ABC) method that combines state-of-the-art dynamical forecasts with observations using machine learning. When applied to the leading subseasonal model from the European Centre for Medium-Range Weather Forecasts (ECMWF), ABC improves temperature forecasting skill by 60-90% and precipitation forecasting skill by 40-69% in the contiguous U.S. We couple these performance improvements with a practical workflow, based on Cohort Shapley, for explaining ABC skill gains and identifying higher-skill windows of opportunity based on specific climate conditions.

1 Introduction

Water and fire managers rely on subseasonal forecasts 2-6 weeks in advance to allocate water, manage wildfires, and prepare for droughts and other weather extremes. However, skillful forecasts for the subseasonal regime are lacking due to the complex dependence on both local weather and global climate variables and the chaotic nature of weather. Bridging the gap between short-term and seasonal forecasting has been the focus of several recent large-scale research efforts which have advanced the subseasonal capabilities of operational physics-based models (1; 2; 3). However, despite these advances, dynamical models still suffer from persistent systematic errors, which limit the skill of temperature and precipitation forecasts for longer lead times from 2 to 6 weeks ahead.

To overcome observed systematic errors of physics-based models on the subseasonal timescale, there have been parallel efforts in recent years to demonstrate the value of machine learning and deep

NeurIPS 2022 AI for Science Workshop.

learning methods in improving subseasonal forecasting (4; 5; 6; 7; 8; 9; 10; 11; 12; 13; 14). While these works demonstrate the promise of statistical models for subseasonal forecasting, they also highlight the complementary strengths of physics- and learning-based approaches and the opportunity to combine those strengths to improve forecasting skill (6; 11).

To harness those complementary strengths, we introduce a hybrid dynamical-learning framework for improved subseasonal forecasting. In particular, we learn to adaptively correct the biases of dynamical models and apply our novel *adaptive bias correction* (ABC) to improve the skill of subseasonal temperature and precipitation forecasts. ABC can be applied operationally as a computationally inexpensive enhancement to any dynamical model forecast, and we use this property to substantially reduce the forecasting errors of eight operational dynamical models, including the state-of-the-art ECMWF model. We couple these performance improvements with a practical workflow for explaining ABC skill gains using Cohort Shapley (15) and identifying higher-skill windows of opportunity (16) based on relevant climate variables. To facilitate future deployment and benchmarking, we release our model and workflow code through the `subseasonal_toolkit` Python package.

2 Methods

We consider two prediction targets: 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). We forecast each variable at $G = 376$ grid points on a $1.5° \times 1.5°$ grid across the contiguous U.S., bounded by latitudes 25N to 50N and longitudes 125W to 67W. To provide the most realistic assessment of forecasting skill (17), all predictions in this study are formed in a real forecast manner that mimics operational use. In particular, to produce a forecast for a given target date, all learning-based models are trained and tuned only on data observable on the corresponding forecast issuance date. We evaluate each forecast according using uncentered anomaly correlation skill. For a collection of target dates, we report average skill using progressive validation (18) to mimic operational use. All data used in this work was obtained from the SubseasonalClimateUSA dataset (19).

2.1 Operational ECMWF and CFSv2 debiasing

We bias correct a uniformly weighted ensemble of the ECMWF control forecast and its 50 ensemble forecasts following the ECMWF operational protocol (20): for each target forecast date, we bias correct the ECMWF 51-member ensemble using the last 20 years of reforecasts with dates within ±6 days from the target month-day combination. The average of ensemble reforecasts on the $1.5° \times 1.5°$ degree grid are used for debiasing.

Following (21), we bias correct the 32-member CFSv2 ensemble forecast in the following way: for each target forecast date, we bias correct the CFSv2 control and ensemble forecasts using the twelve-year period from 1999 to 2010 of reforecasts. The average of the ensemble reforecasts on the $1.5° \times 1.5°$ degree grid are used for debiasing.

2.2 Adaptive bias correction

ABC is a uniformly-weighted ensemble of three machine learning models that we introduce in this work and describe below in details: Climatology++, Dynamical++, and Persistence++. Supplementary algorithm descriptions can be found in Appendix B.

After averaging dynamical forecasts over a range of issuance dates and lead times, **Dynamical++** debiases the ensemble forecast for each grid cell by adding the mean value of the target variable and subtracting the mean forecast over a learned window of observations around the target day of year. For a given target date $t^\star$ and lead time $l^\star$, the Dynamical++ training set $\mathcal{T}$ is restricted to data fully observable one day prior to the issuance date, that is, to dates $t \leq t^\star - l^\star - L - 1$ where $L = 14$ represents the forecast period length. Unlike standard debiasing strategies, which employ static ensembling and bias correction, Dynamical++ adaptively selects the range of ensembled lead times $\mathcal{L}$, the number of averaged issuance dates $d^\star$, and the size s of the observation window using an automated tuning procedure.

For each target date, Dynamical++ is run with the hyperparameter configuration that achieved the smallest mean progressive geographic root mean squared error (RMSE) over the preceding 3 years. Here, *progressive* indicates that each candidate model forecast is generated using all training data observable prior to the associated forecast issuance date. Every configuration with $s \in \{0, 14, 28, 35\}$, $d^\star \in \{1, 7, 14, 28, 42\}$, and $\mathcal{L} = \{29\}$ for the weeks 5-6 lead time and $\mathcal{L} \in \{\{15\}, [15, 22], [0, 29], \{29\}\}$ the weeks 3-4 lead time was considered.

Climatology++ is a locally constant prediction rule that minimizes historical forecasting error, specified by a user-supplied loss function, over all days in a window around the target day of year. For a given target date $t^\star$ and lead time $l^\star$, the Climatology++ training set $\mathcal{T}$ is restricted to data fully observable one day prior to the issuance date, that is, to dates $t \le t^\star - l^\star - L - 1$ where $L = 14$ represents the forecast period length. The number of training years Y and the size of the observation window (quantified by the *span*, the number of days s included on each side of the target day of year) are determined adaptively using an automated tuning procedure, described below.

For each target date, Climatology++ is run with the hyperparameter configuration that achieved the smallest mean progressive geographic RMSE over the preceding 3 years. All spans $s \in \{0, 1, 7, 10\}$ were considered. All precipitation configurations used the geographic MSE loss and all available training years. All temperature configurations used the geographic RMSE loss and either all available training years or $Y = 29$.

Persistence++ fits a least-squares regression per grid point to optimally combine lagged temperature or precipitation measurements, climatology, and a dynamical ensemble forecast. For a given target date $t^\star$ and lead time $l^\star$, the Persistence++ training set $\mathcal{T}$ is restricted to data fully observable one day prior to the issuance date, that is, to dates $t \le t^\star - l^\star - L - 1$ where $L = 14$ represents the forecast period length.

3 Results

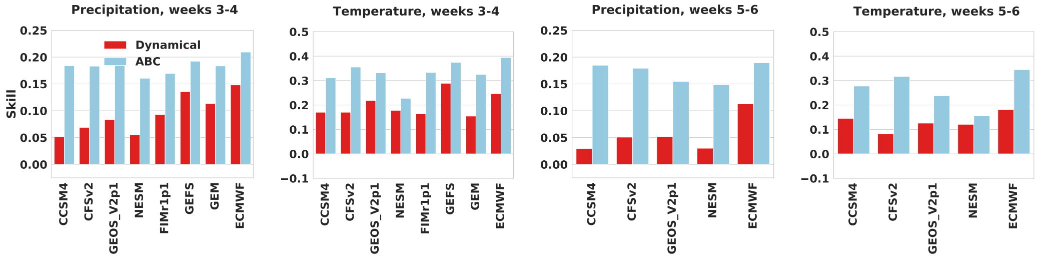

Figure 1: Average model skill for ECWMF and SubX dynamical models (red) and their ABC-corrected counterparts (blue) across the contiguous U.S. and the years 2018–2021. For each forecasting task and dynamical model input, ABC provides a pronounced improvement in skill.

Figure 1 highlights the advantage of ABC over raw dynamical models when forecasting accumulated precipitation and averaged temperature in the contiguous U.S. Here, ABC is applied to the leading subseasonal model, ECMWF, and to each of seven operational models participating in the Subseasonal Experiment (SubX, 2). Subseasonal forecasting skill, measured by uncentered anomaly correlation, is evaluated at two forecast horizons, weeks 3-4 and weeks 5-6, and averaged over all available forecast dates in the four-year span 2018–2021. We find that, for each dynamical model input and forecasting task, ABC leads to a pronounced improvement in skill. For example, when applied to the U.S. operational model CFSv2, ABC improves temperature forecasting skill by 109-289% and precipitation skill by 165-253%. When applied to the leading ECMWF model, ABC improves temperature skill by 60-90% and precipitation skill by 40-69%. Moreover, for precipitation, even lower-skill models like CCSM4 enjoy skill comparable to the best after the application of ABC. Overall and despite significant variability in dynamical model skill, ABC consistently reduces the systematic errors of its input model, bringing forecasts closer to observations for each target variable and time horizon.

We next examine, in Figure 2, the spatial distribution of skill for CFSv2, ECMWF, and their ABC-corrected counterparts at three forecast horizons. At the shorter-term horizon of weeks 1-2, both CFSv2 and ECMWF enjoy reasonably high skill throughout the contiguous U.S. However, skill drops

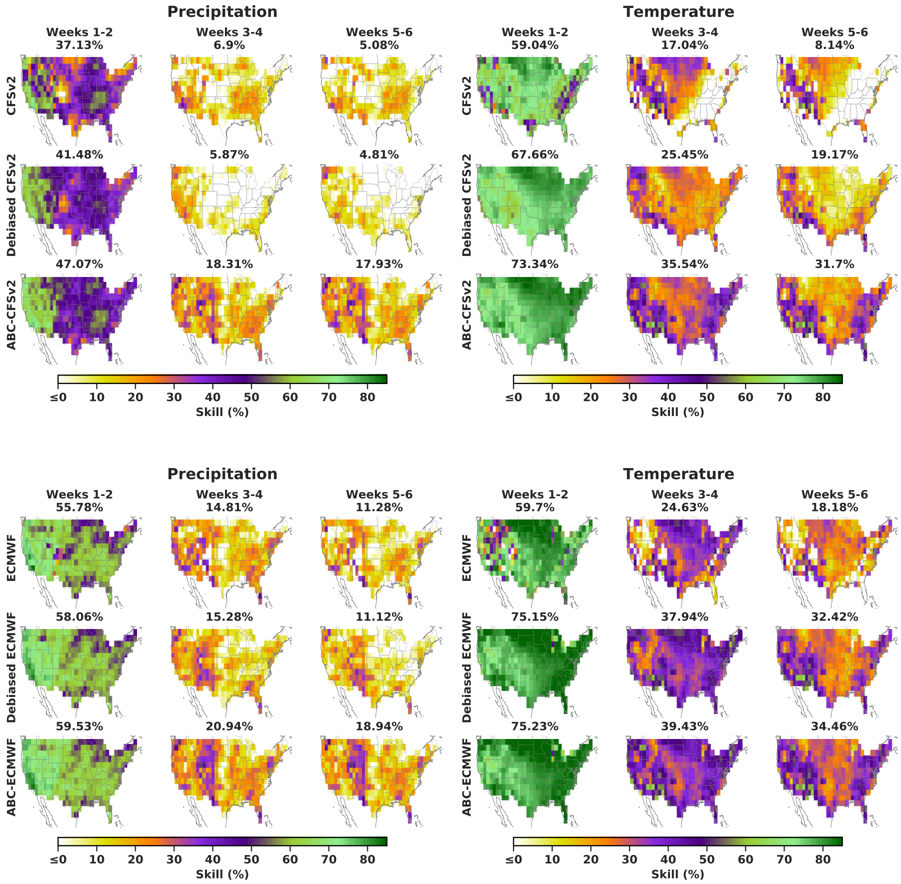

Figure 2: Spatial skill distribution of dynamical models and ABC corrections. Across the contiguous U.S. and the years 2018–2021, dynamical model skill drops precipitously at subseasonal timescales (weeks 3-4 and 5-6), but ABC attenuates the degradation, doubling or tripling the skill of CFSv2 and boosting ECMWF skill 40-90%. Taking the same raw model forecasts as input, ABC also provides consistent improvements over operational debiasing protocols, tripling the precipitation skill of debiased CFSv2 and improving that of debiased ECMWF by 70%. The average skill over all sites is displayed above each map.

precipitously for both models when moving to the subseasonal horizons (weeks 3-4 and 5-6). This degradation is particularly striking for precipitation, where prediction skill drops to zero or below in the central and northeastern parts of the U.S. For temperature prediction, CFSv2 has a skill of zero across a broad region of the East, while ECMWF produces isolated pockets of zero skill in the West. At these subseasonal timescales, ABC provides consistent improvements across the U.S. that either double or triple the mean skill of CFSv2 and increase the mean skill of ECMWF by 40-90%. Notably, ABC also improves over standard operational debiasing protocols (labeled debiased CFSv2 and debiased ECMWF in Figure 2), tripling the average precipitation skill of debiased CFSv2 and increasing that of debiased ECMWF by 70%.

An important component contributing to the overall accuracy of ABC is a reduction of the systematic bias introduced by its dynamical model input. Figure 3 examines the spatial distribution of this bias by plotting the average difference between forecasts and observations over all forecast dates. The precipitation maps reveal a wet bias over the northern half of the U.S. for CFSv2 (average bias: 8.32 mm) and a dry bias over the south-east part of the U.S. for ECMWF (average bias: −8.12 mm). In this case, ABC eliminates the CFSv2 wet bias (average bias: −0.46 mm) and slightly alleviates the ECMWF dry bias (average bias: −6.24 mm). For temperature, we observe a cold bias over the eastern half of the U.S. for CFSv2 (average bias: −1.2°C) and notice a mixed pattern of cold and warm biases over the western half of the U.S for ECMWF (average bias: −0.30°C). In this case, although ABC does not eliminate these biases entirely, it reduces the magnitude of the cold eastern bias by bringing CFSv2 forecasts closer to observations (average bias: −0.18C) and reduces the mixed ECMWF bias (average bias: −0.04C).

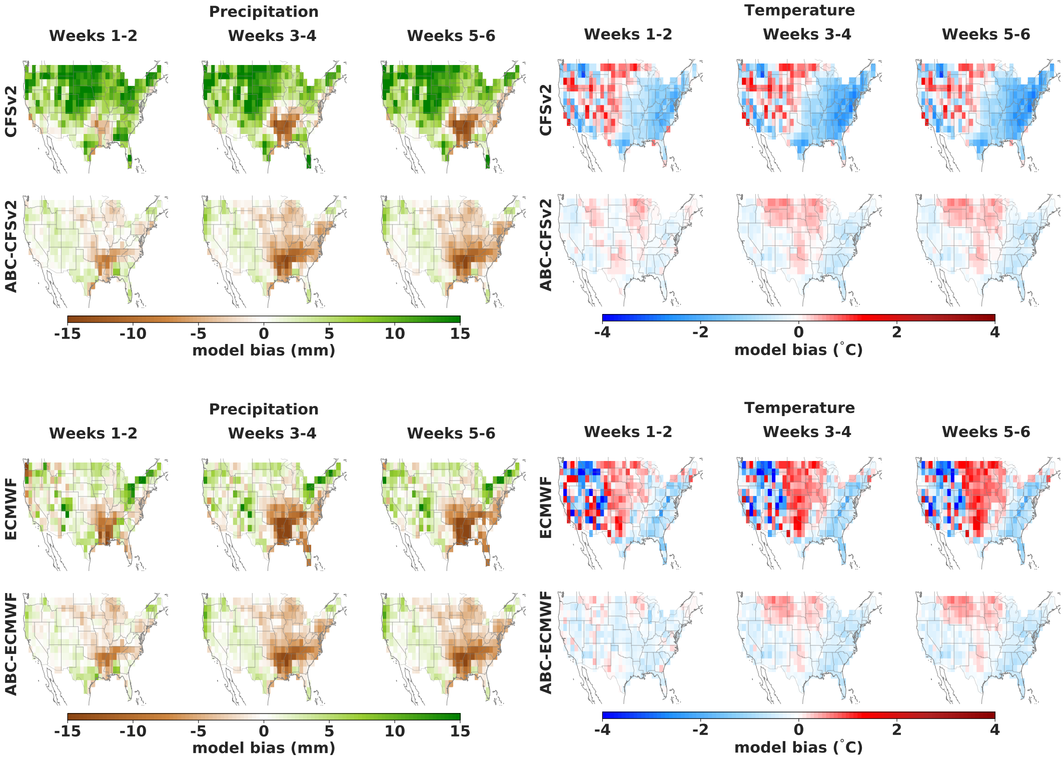

Figure 3: Spatial distribution of model bias over the years 2018–2021. Across the contiguous U.S., ABC reduces the systematic model bias of its dynamical model input for both temperature and precipitation.

The results presented so far assess overall model skill, averaged across all forecast dates. However, there is a growing appreciation that subseasonal forecasts can also benefit from selective deployment during "windows of opportunity," periods defined by observable climate conditions in which specific forecasters are likely to have higher skill (16). Here, we propose a practical *opportunistic ABC workflow* that uses a candidate set of explanatory variables to identify windows in which ABC is especially likely to improve upon a baseline model. This workflow is described in details in Appendix A. The same workflow can be used to explain the skill improvements achieved by ABC in terms of the explanatory variables.

The opportunistic ABC workflow is based on the optimal credit assignment principle (22) and measures the impact of explanatory variables on individual forecasts using Cohort Shapley (15) and overall variable importance using Shapley effects (23). We use these Shapley measures to interpret the contexts in which ABC offers improvements in terms of climate variables with known relevance for subseasonal forecasting skill. As a running example, we use our workflow to explain the skill differences between ABC-ECMWF and debiased ECMWF when predicting precipitation in weeks 3-4. As our candidate explanatory variables we use Northern Hemisphere geopotential heights (HGT) at 500 and 10 hPa, the phase of the Madden-Julian Oscillation (MJO), Northern Hemisphere sea ice concentration (ICEC), global sea surface temperatures (SST), the multivariate El Niño-Southern Oscillation index (MEI.v2, 24), and the target month. All variables are lagged appropriately to ensure that they are observable on the forecast issuance date.

We first use Shapley effects to determine the overall importance of each variable in explaining the precipitation skill improvements of ABC-ECMWF. We find the most important explanatory variables to be the first two principal components (PCs) of 500 hPa geopotential height, the MJO phase, the second PC of 10 hPa geopotential height, and the first PC of sea ice concentration. These

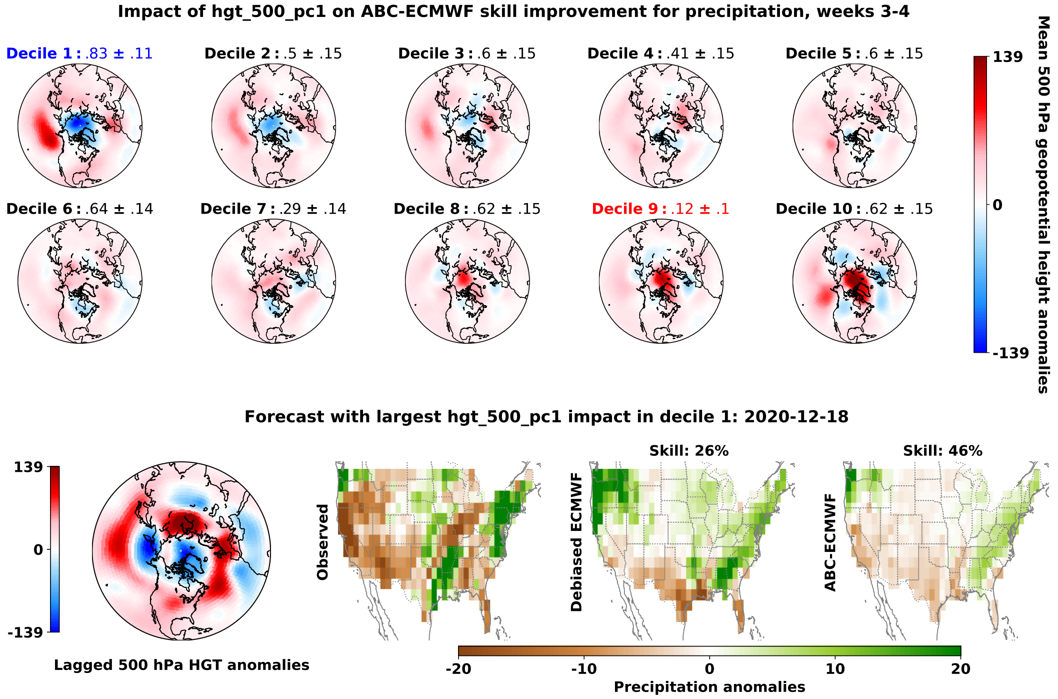

Figure 4: **Top:** To summarize the impact of `hgt_500_pc1`on ABC-ECMWF skill improvement for precipitation weeks 3-4, we divide our forecasts into 10 bins, determined by the deciles of `hgt_500_pc1`, and compute the probability of positive impact in each bin, as shown above each bin map. The highest probabilities of positive impact are shown in blue and the lowest probabilities of positive impact are shown in red. We find that `hgt_500_pc1`is most likely to have a positive impact on skill improvement in decile 1, which features a positive Arctic Oscillation (AO) pattern, and least likely in decile 9, which features AO in the opposite phase. **Bottom:** The forecast most impacted by `hgt_500_pc1`in decile 1 is also preceded by a positive AO pattern and replaces the wet debiased ECMWF forecast with a more skillful dry pattern in the west.

variables are consistent with the literature exploring the dominant contributions to subseasonal precipitation (25; 26; 27; 28).

We next use Cohort Shapley to identify the contexts in which each variable has the greatest impact on skill. For example, Figure 4 summarizes the impact of the first 500 hPa geopotential heights PC (`hgt_500_pc1`) on ABC-ECMWF skill improvement. This display divides our forecasts into 10 bins, determined by the deciles of `hgt_500_pc1`, and computes the probability of positive impact in each bin. We find that `hgt_500_pc1` is most likely to have a positive impact impact on skill improvement in decile 1, which features a positive Arctic Oscillation (AO) pattern, and least likely in decile 9, which features AO in the opposite phase. The ABC-ECMWF forecast most impacted by `hgt_500_pc1` in decile 1 is also preceded by a positive AO pattern and replaces the wet debiased ECMWF forecast with a more skillful dry pattern in the west.

Finally, we use the identified contexts to define windows of opportunity for operational deployment. Indeed, since all explanatory variables are observable on the forecast issuance date, one can selectively apply ABC when multiple variables are likely to have a positive impact on skill and otherwise issue a default, standard forecast (e.g., debiased ECMWF). We call this selective forecasting model *opportunistic ABC*. How many high-impact variables should we require when defining these windows of opportunity? Requiring a larger number of high-impact variables will tend to increase the skill gains of ABC but simultaneously reduce the number of dates on which ABC is deployed. Figure 5 illustrates this trade-off for ABC-ECMWF and shows that opportunistic ABC skill is maximized when two or more high-impact variables are required. With this choice, ABC is used for approximately 81% of forecasts and debiased ECMWF is used for the remainder.

# High-impact	% Forecasts	High-impact skill (%)	
variables	using ABC	ABC	Debiased
0 or more	100.00	20.94	15.28
1 or more	95.93	20.99	14.84
2 or more	80.62	22.29	13.12
3 or more	58.61	23.56	12.00
4 or more	31.82	24.72	8.18
5 or more	14.59	26.51	8.35
6 or more	6.46	29.72	10.55
7 or more	2.15	45.00	17.53

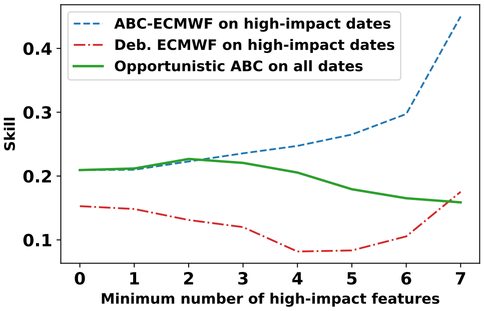

Figure 5: Defining windows of opportunity for opportunistic ABC forecasting of precipitation weeks 3-4. **Left:** When more explanatory variables fall into high-impact deciles or bins (e.g., the blue bins of Figure 4), the mean skill of ABC-ECMWF improves, but the percentage of forecasts using ABC declines. **Right:** The overall skill of opportunistic ABC is maximized when ABC-ECMWF is deployed for target dates with two or more high-impact variables and standard debiased ECMWF is deployed otherwise.

4 Discussion

Dynamical models have shown increasing skill in accurately forecasting the weather (29), but they still contain systematic biases that compound on subseasonal time scales and suppress forecast skill (30; 31; 32; 33). Our proposed solution, ABC, learns to correct these biases by adaptively integrating dynamical forecasts, historical observations, and recent weather trends. When applied to the leading subseasonal model from ECMWF, ABC improves forecast skill by 60-90% for precipitation and 40-69% for temperature. The same approach substantially reduces the forecasting errors of seven additional operational models, with less skillful input models performing nearly as well as ECMWF after correction. This finding suggests that systematic errors in dynamical models are a primary contributor to observed skill differences and that ABC provides an effective mechanism for reducing these heterogeneous errors. Because ABC is also simple to implement and deploy in real-time operational settings, adaptive bias correction represents a computationally inexpensive strategy for upgrading operational models, while conserving valuable human resources.

Moreover, ABC is, by its nature, adaptive to changes in systematic biases. As operational models are upgraded and systematic biases evolve, our ABC training protocol is designed to ingest the upgraded model forecasts and hindcasts reflecting those changes. In addition, the same protocol can be adapted to correct probabilistic subseasonal forecasts that estimate the distribution of future weather (34).

To capitalize on higher-skill forecasts of opportunity, we have also introduced an opportunistic ABC workflow that explains the skill improvements of ABC in terms of a candidate set of environmental variables, identifies high-probability windows of opportunity based on those variables, and selectively deploys either ABC or a baseline forecast to maximize expected skill. The same workflow can be applied to explain the skill improvements of any forecasting model and, unlike other popular explanation tools (e.g., 35; 36), requires no expensive model retraining, no generation of additional forecasts beyond those routinely generated for operational or hindcast use, and allows for explanations in terms of variables that were not explicitly used in training the model.

Overall, we find that correcting dynamical forecasts using ABC yields an effective and scalable strategy for building the next generation of subseasonal forecasting models. We anticipate that our hybrid dynamical-learning framework will benefit both research and operations, and we release our open-source code to facilitate future adoption and development.

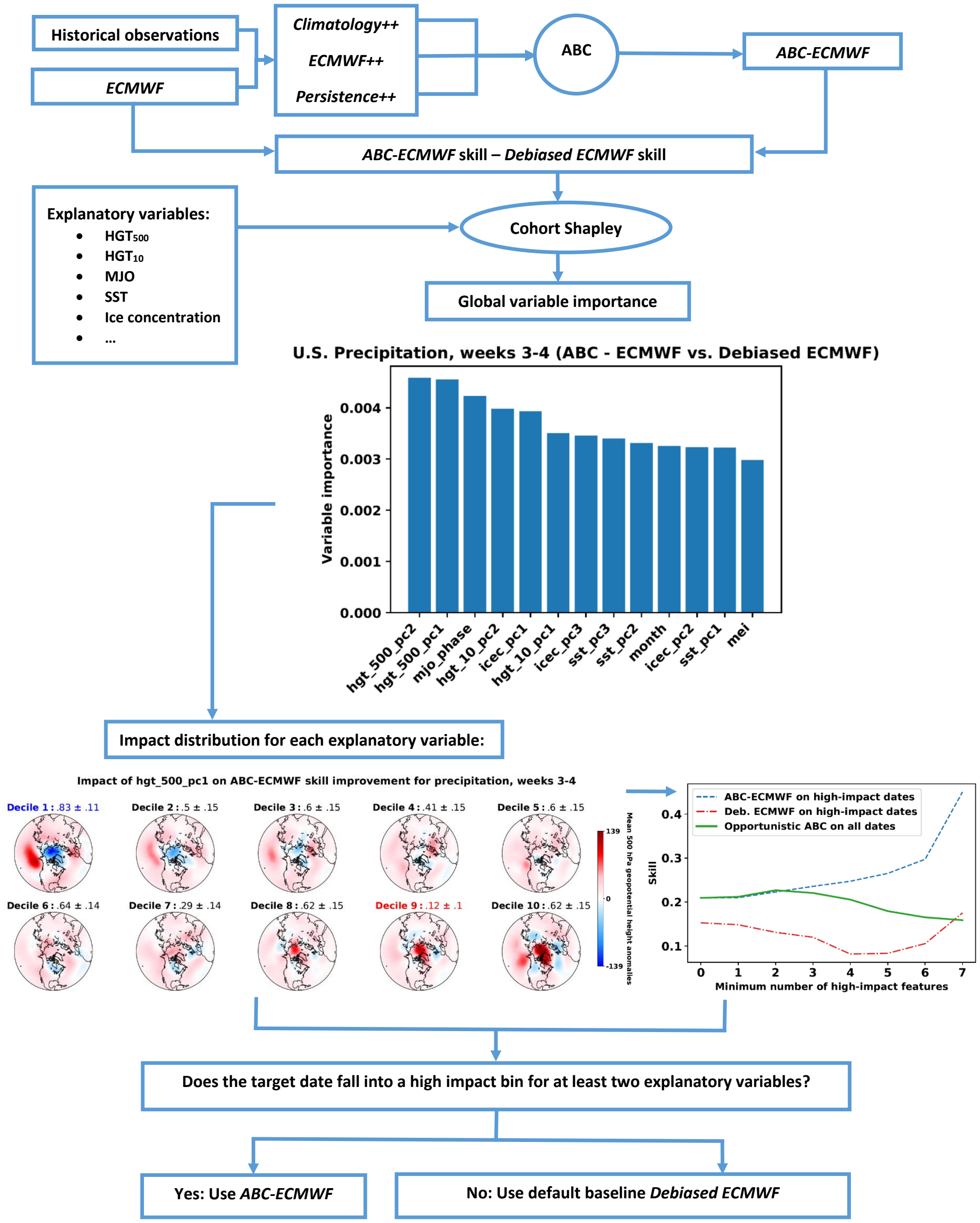
Historical observations
ECMWF
Climatology++
ECMWF++
Persistence++
ABC
ABC-ECMWF
ABC-ECMWF skill – Debiased ECMWF skill
Explanatory variables:
HGT500
HGT10
MJO
SST
Ice concentration
...
Cohort Shapley
Global variable importance
U.S. Precipitation, weeks 3-4 (ABC - ECMWF vs. Debiased ECMWF)
Variable importance
0.000
0.001
0.002
0.003
0.004
hgt_500_pc2
hgt_500_pc1
mjo_phase
hgt_10_pc2
icec_pc1
hgt_10_pc1
icec_pc3
sst_pc3
sst_pc2
month
icec_pc2
sst_pc1
mei
Impact distribution for each explanatory variable:
Impact of hgt_500_pc1 on ABC-ECMWF skill improvement for precipitation, weeks 3-4
Decile 1 : .83 ± .11
Decile 2 : .5 ± .15
Decile 3 : .6 ± .15
Decile 4 : .41 ± .15
Decile 5 : .6 ± .15
Decile 6 : .64 ± .14
Decile 7 : .29 ± .14
Decile 8 : .62 ± .15
Decile 9 : .12 ± .1
Decile 10 : .62 ± .15
139
0
-139
Mean 500 hPa geopotential height anomalies
ABC-ECMWF on high-impact dates
Deb. ECMWF on high-impact dates
Opportunistic ABC on all dates
Skill
0.1
0.2
0.3
0.4
0
1
2
3
4
5
6
7
Minimum number of high-impact features
Does the target date fall into a high impact bin for at least two explanatory variables?
Yes: Use ABC-ECMWF
No: Use default baseline Debiased ECMWF

Figure 6: Schematic of the opportunistic ABC workflow. Opportunistic ABC uses historical ABC and baseline forecasts and a candidate set of explanatory variables to identify windows of opportunity for selective deployment of ABC in an operational setting.

References

[1] 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*, vol. 98, no. 1, pp. 163–173, 2017.

[2] K. Pegion, B. P. Kirtman, E. Becker, D. C. Collins, E. LaJoie, R. Burgman, R. Bell, T. DelSole, D. Min, Y. Zhu, *et al.*, "The subseasonal experiment (subx): A multimodel subseasonal prediction experiment," *Bulletin of the American Meteorological Society*, vol. 100, no. 10, pp. 2043–2060, 2019.

[3] A. L. Lang, K. Pegion, and E. A. Barnes, "Introduction to special collection:"bridging weather and climate: Subseasonal-to-seasonal (S2S) prediction"," *Journal of Geophysical Research: Atmospheres*, vol. 125, no. 4, p. e2019JD031833, 2020.

[4] L. Li, R. W. Schmitt, C. C. Ummenhofer, and K. B. Karnauskas, "Implications of north atlantic sea surface salinity for summer precipitation over the us midwest: Mechanisms and predictive value," *Journal of Climate*, vol. 29, no. 9, pp. 3143–3159, 2016.

[5] 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.

[6] J. Hwang, P. Orenstein, J. Cohen, K. Pfeiffer, and L. Mackey, "Improving subseasonal forecasting in the western U.S. with machine learning," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '19, (New York, NY, USA), p. 2325–2335, Association for Computing Machinery, 2019.

[7] 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.

[8] 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.

[9] 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.

[10] C. Wang, Z. Jia, Z. Yin, F. Liu, G. Lu, and J. Zheng, "Improving the accuracy of subseasonal forecasting of china precipitation with a machine learning approach. front," *Earth Sci*, vol. 9, p. 659310, 2021.

[11] M. Kim, C. Yoo, and J. Choi, "Enhancing subseasonal temperature prediction by bridging a statistical model with dynamical arctic oscillation forecasting," *Geophysical Research Letters*, vol. 48, no. 15, p. e2021GL093447, 2021.

[12] 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.

[13] 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," *Journal of Advances in Modeling Earth Systems*, vol. 13, no. 7, p. e2021MS002502, 2021.

[14] V. Srinivasan, J. Khim, A. Banerjee, and P. Ravikumar, "Subseasonal climate prediction in the western us using bayesian spatial models," in *Uncertainty in artificial intelligence*, vol. 37, 2021.

[15] M. Mase, A. B. Owen, and B. Seiler, "Explaining black box decisions by shapley cohort refinement," *arXiv preprint arXiv:1911.00467*, 2019.

[16] A. Mariotti, C. Baggett, E. A. Barnes, E. Becker, A. Butler, D. C. Collins, P. A. Dirmeyer, L. Ferranti, N. C. Johnson, J. Jones, *et al.*, "Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond," *Bulletin of the American Meteorological Society*, vol. 101, no. 5, pp. E608–E625, 2020.

[17] J. S. Risbey, D. T. Squire, A. S. Black, T. DelSole, C. Lepore, R. J. Matear, D. P. Monselesan, T. S. Moore, D. Richardson, A. Schepen, *et al.*, "Standard assessments of climate forecast skill can be misleading," *Nature Communications*, vol. 12, no. 1, pp. 1–14, 2021.

[18] A. Blum, A. Kalai, and J. Langford, "Beating the hold-out: Bounds for k-fold and progressive cross-validation," in *Proceedings of the twelfth annual conference on Computational learning theory*, pp. 203–208, 1999.

[19] S. dataset, “Subseasonal data Python package.” `https://github.com/microsoft/subseasonal_data`, 2021.

[20] ECMWF, “Re-forecast for medium and extended forecast range.” `https://www.ecmwf.int/en/forecasts/documentation-and-support/extended-range/re-forecast-medium-and-extended-forecast-range`, 2022. Accessed: 2022-06-29.

[21] K. Nowak, R. Webb, R. Cifelli, and L. Brekke, “Sub-seasonal climate forecast rodeo,” in *2017 AGU Fall Meeting, New Orleans, LA*, pp. 11–15, 2017.

[22] L. Shapley, “A value for n-person games. contributions to the theory of games ii, kuhn, h., tucker, a,” 1953.

[23] E. Song, B. L. Nelson, and J. Staum, “Shapley effects for global sensitivity analysis: Theory and computation,” *SIAM/ASA Journal on Uncertainty Quantification*, vol. 4, no. 1, pp. 1060–1083, 2016.

[24] K. Wolter and M. S. Timlin, “Monitoring enso in coads with a seasonally adjusted principal,” in *Proc. of the 17th Climate Diagnostics Workshop, Norman, OK, NOAA/NMC/CAC, NSSL, Oklahoma Clim. Survey, CIMMS and the School of Meteor., Univ. of Oklahoma, 52*, vol. 57, 1993.

[25] N. Christidis and P. A. Stott, “Changes in the geopotential height at 500 hpa under the influence of external climatic forcings,” *Geophysical Research Letters*, vol. 42, no. 24, pp. 10–798, 2015.

[26] S. J. Woolnough, “Chapter 5 - the madden-julian oscillation,” in *Sub-Seasonal to Seasonal Prediction* (A. W. Robertson and F. Vitart, eds.), pp. 93–117, Elsevier, 2019.

[27] W. J. Merryfield, J. Baehr, L. Batté, E. J. Becker, A. H. Butler, C. A. Coelho, G. Danabasoglu, P. A. Dirmeyer, F. J. Doblas-Reyes, D. I. Domeisen, *et al.*, “Current and emerging developments in subseasonal to decadal prediction,” *Bulletin of the American Meteorological Society*, vol. 101, no. 6, pp. E869–E896, 2020.

[28] M. Chevallier, F. Massonnet, H. Goessling, V. Guémas, and T. Jung, “Chapter 10 - the role of sea ice in sub-seasonal predictability,” in *Sub-Seasonal to Seasonal Prediction* (A. W. Robertson and F. Vitart, eds.), pp. 201–221, Elsevier, 2019.

[29] P. Bauer, A. Thorpe, and G. Brunet, “The quiet revolution of numerical weather prediction,” *Nature*, vol. 525, no. 7567, pp. 47–55, 2015.

[30] I. P. on Climate Change, *Evaluation of Climate Models*, p. 741–866. Cambridge University Press, 2014.

[31] A. Zadra, K. Williams, A. Frassoni, M. Rixen, Ángel F. Adames, J. Berner, F. Bouyssel, B. Casati, H. Christensen, M. B. Ek, G. Flato, Y. Huang, F. Judt, H. Lin, E. Maloney, W. Merryfield, A. V. Niekerk, T. Rackow, K. Saito, N. Wedi, and P. Yadav, “Systematic errors in weather and climate models: Nature, origins, and ways forward,” *Bulletin of the American Meteorological Society*, vol. 99, no. 4, pp. ES67 – ES70, 2018.

[32] L. Zhang, T. Kim, T. Yang, Y. Hong, and Q. Zhu, “Evaluation of subseasonal-to-seasonal (S2S) precipitation forecast from the north american multi-model ensemble phase ii (nmme-2) over the contiguous us,” *Journal of Hydrology*, vol. 603, p. 127058, 2021.

[33] E. Dutra, F. Johannsen, and L. Magnusson, “Late spring and summer subseasonal forecasts in the northern hemisphere midlatitudes: Biases and skill in the ecmwf model,” *Monthly Weather Review*, vol. 149, no. 8, pp. 2659–2671, 2021.

[34] T. Palmer, “The primacy of doubt: Evolution of numerical weather prediction from determinism to probability,” *Journal of Advances in Modeling Earth Systems*, vol. 9, no. 2, pp. 730–734, 2017.

[35] M. T. Ribeiro, S. Singh, and C. Guestrin, “"Why should I trust you?" explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[36] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in neural information processing systems*, vol. 30, 2017.

A Supplementary Method Details

Here we detail the steps of the opportunistic ABC workflow illustrated in Figure 6 using ECMWF as an example dynamical input. The same workflow applies to any other dynamical input.

1. Identify a set of V candidate explanatory variables. Here we use the temporal variables enumerated in (6, Fig. 2) augmented with the first two PCs of 500 hPa geopotential heights and the target month. To ensure that the workflow can be deployed operationally, we use lagged observations with lags chosen so that each variable is observable on the forecast issuance date.
2. Compute the skill difference between ABC-ECMWF and debiased ECMWF for each target date in the evaluation period.
3. Use the `cohortshapley` Python package to compute global variable importances (measured by Shapley effects) and forecast-specific variable impact values explaining the skill differences.
4. For each continuous explanatory variable (e.g., hgt_500_pc2), divide the evaluation period forecasts into 10 bins, determined by the deciles of the explanatory variable. For each categorical variable (e.g., mjo_phase), divide the forecasts into bins determined by the categories (e.g., MJO phases).
5. Estimate the probability of positive variable impact in each bin and compute a 95% bootstrap confidence interval. Flag all bin probabilities within the confidence interval of the highest probability bin as high impact and all bin probabilities within the confidence interval of the lowest probability bin as low impact. Visualize and interpret the highest and lowest impact bins.
6. Identify the forecast most impacted by the explanatory variable in the high impact bins. Visualize the ABC-ECMWF and debiased ECMWF forecasts and the associated explanatory variable for that target date.
7. For each $k \in \{0, \ldots, V\}$, compute opportunistic ABC skill when k or more explanatory variables fall into high impact bins. Let $k^\star$ represent the integer at which opportunistic ABC skill is maximized.
8. At each future forecast issuance date, deploy ABC-ECMWF if $k^\star$ or more explanatory variables fall into high impact bins and deploy debiased ECMWF otherwise.

B Supplementary Model Details

This section presents the algorithm details for the three machine learning models underlying ABC: Dynamical++, Climatology++, and Persistence++.

Algorithm 1 Dynamical++

```
input test date $t^\star$; lead time $l^\star$; # issuance dates $d^\star$; span $s$; training set ground truth and dynamical forecasts $(\mathbf{y}_t, \mathbf{f}_{t,l})_{t\in\mathcal{T}, l\in\mathcal{L}}$
initialize days per year $D = 365.242199$; # training years $Y = 12$
  $\mathcal{S} = \{t \in \mathcal{T} : \texttt{year\_diff} := \lfloor \frac{t^\star - t}{D} \rfloor \leq Y$ and $\texttt{day\_diff} := \frac{365}{2} - |\lfloor (t^\star - t) \mod D \rfloor - \frac{365}{2}| \leq s\}$
  // Form dynamical ensemble forecast across issuance dates and lead times $l \in \mathcal{L}$
  for training and test dates $t \in \mathcal{S} \cup \{t^\star\}$ do
    $\bar{\mathbf{f}}_t = \text{mean}((\mathbf{f}_{t-l^\star-d+1,l})_{1\leq d\leq d^\star, l\in\mathcal{L}})$
output $\bar{\mathbf{f}}_{t^\star} + \text{mean}((\mathbf{y}_t - \bar{\mathbf{f}}_t)_{t\in\mathcal{S}})$
```

Algorithm 2 Climatology++

```
input test date $t^\star$; # training years $Y$; span $s$; loss $\in \{\text{RMSE},\text{MSE}\}$; training set ground truth $(\mathbf{y}_t)_{t\in\mathcal{T}}$
initialize days per year $D = 365.242199$
  $\mathcal{S} = \{t \in \mathcal{T} : \texttt{year\_diff} := \lfloor \frac{t^\star - t}{D} \rfloor \leq Y$ and $\texttt{day\_diff} := \frac{365}{2} - |\lfloor (t^\star - t) \mod D \rfloor - \frac{365}{2}| \leq s\}$
output $\text{argmin}_{\mathbf{y}} \sum_{t\in\mathcal{S}} \text{loss}(\mathbf{y}, \mathbf{y}_t)$
```

Algorithm 3 Persistence++

```
input lead time $l^\star$; training set ground truth, climatology, and dynamical forecasts $(\mathbf{y}_t, \mathbf{c}_t, \mathbf{f}_{t,l})_{t\in\mathcal{T}, l\in\mathcal{L}}$
initialize forecast period length $L = 14$
  // Form dynamical ensemble forecast across subseasonal lead times $l \geq l^\star$
  for training dates $t \in \mathcal{T}$ do
    $\bar{\mathbf{f}}_t = \text{mean}((\mathbf{f}_{t,l})_{l\geq l^\star})$
  // Combine ensemble forecast, climatology, and lagged measurements
  for grid points $g = 1$ to $G$ do
    $\hat{\boldsymbol{\beta}}_g \in \text{argmin}_{\boldsymbol{\beta}} \sum_{t\in\mathcal{T}} (y_{t,g} - \boldsymbol{\beta}^\top [1, c_{t,g}, y_{t-l^\star-L-1,g}, y_{t-2l^\star-L-1,g}, \bar{f}_{t-l^\star-1,g}])^2$
output coefficients $(\hat{\boldsymbol{\beta}}_g)_{g=1}^G$
```