Bridging the "Predictability Desert": A Probabilistic Bias Correction Framework for AI and Dynamical Subseasonal Forecasts

Hannah Guan Harvard College
Soukayna Mouatadid University of Toronto
Paulo Orenstein IMPA
Judah Cohen MIT
Haiyu Dong Microsoft
Genevieve Flaspohler Rhiza Research
Alex Lu Microsoft Research
Jonathan Weyn Microsoft
Lester Mackey Microsoft Research

Abstract

Decision-makers rely on weather forecasts to allocate water resources, manage wildfires, plant crops, and prepare for weather extremes. Today, such forecasts enjoy unprecedented accuracy out to two weeks thanks to steady advances in physics-based dynamical models and data-based artificial intelligence (AI) models. However, model skill drops precipitously at subseasonal timescales (weeks 3 – 6) due to compounding errors and persistent biases. To counter this degradation, we introduce *probabilistic bias correction (PBC)*, a machine learning framework that substantially reduces systematic error by learning to correct historical probabilistic forecasts. When applied to the leading dynamical model from the European Centre for Medium-Range Weather Forecasts (ECMWF), PBC boosts subseasonal forecasting skill by 70–80% for precipitation and over 200% for temperature and sea-level pressure. We designed PBC for operational deployment, and, in ECMWF's 2025 real-time forecasting competition, its global forecasts placed first for all weather variables and lead times, outperforming the dynamical models of six government agencies, ECMWF's AI forecasting system, and the forecasting systems of 34 teams worldwide. These probabilistic skill gains translate into more accurate prediction of extreme events and have the potential to improve agricultural planning, energy management, and disaster preparedness.

Workshop "Towards Trustworthy Predictions: Theory and Applications of Calibration for Modern AI" at AISTATS 2026, Tangier, Morocco.

1 Introduction

Decision-makers around the world rely on subseasonal forecasts—weather predictions 3 to 6 weeks ahead—to allocate water resources, manage wildfires, make planting decisions, and prepare for natural disasters (Merryfield et al., 2020). Yet this timescale remains notoriously difficult to predict. While modern dynamical and AI-based forecasting systems achieve high skill at short lead times, performance degrades rapidly beyond two weeks due to compounding model errors and the chaotic nature of the weather (Lorenz, 1969; Nathaniel et al., 2024), and even state-of-the-art systems often struggle to outperform simple climatological baselines at subseasonal ranges (Chen et al., 2024).

A core difficulty is that leading subseasonal models are generated iteratively over many time steps, and even small per-step modeling errors compound into large systematic errors over several weeks' time (Nathaniel et al., 2024). Indeed, subseasonal forecasting has long been considered a "predictability desert" due to compounding model errors and complex dependencies on both local weather and global climate variables (Richter and Joseph, 2025).

Here we show that while existing subseasonal ensembles are indeed error-prone, their systematic errors are both predictable and correctable using machine learning, allowing for substantial gains in skill with minimal cost. To achieve this, we introduce a *probabilistic bias correction (PBC)* framework that learns to correct a model's forecast distribution using its own historical predictions and observations of past weather. PBC is compatible with any input model, and we use it to advance the state of the art in both data-driven and hybrid (AI + dynamical) subseasonal forecasting. The resulting forecasts are designed for operational deployment and, in real-time competition, outperform the best alternative dynamical, AI, and hybrid forecasting systems from around the globe.

2 Methods

We use the ERA5 reanalysis dataset as ground truth for training and evaluation. ERA5 provides hourly estimates of atmospheric variables on a global $0.25° \times 0.25°$ grid. We aggregate ERA5 data to weekly averages and regrid to a $1.5° \times 1.5°$ resolution. The target variables are: 2-meter temperature (tas), total precipitation (pr), and mean sea level pressure (mslp).

We formulate subseasonal prediction as a probabilistic forecasting task over climatological quintiles. For each variable, grid point, and calendar day, we compute quintile thresholds from a 20-year rolling climatology of ERA5 observations, computed with a ± 4 day smoothing window around each calendar day. These ERA5-based quintile thresholds define the target categories for probabilistic forecasting and are used consistently throughout our framework for both training and evaluation.

To generate probabilistic forecasts from ensemble predictions, we convert ensemble forecasts into cumulative probabilities at the ERA5-based quintile thresholds. For each target date, grid point, and quintile threshold, the raw CDF forecast is computed as the fraction of ensemble members predicting values at or below the threshold. This formulation directly uses the observation-based climatology to define forecast categories.

2.1 Probabilistic Bias Correction (PBC)

Probabilistic bias correction (PBC) is a new machine learning framework for improving subseasonal forecasts from dynamical, AI, or hybrid prediction systems. Given a raw ensemble of deterministic forecasts, PBC outputs learning-enhanced probabilities designed to correct systematic errors while preserving the predictive signal in the underlying model. Our approach is inspired by the deterministic bias correction framework of Mouatadid et al. (2023) but operates directly on the space of probability distributions rather than at the level of raw observations. The added flexibility allows us to directly optimize not only the center of a distribution but also its spread and shape.

Figure 1 illustrates an example application of PBC. First, an input ensemble is converted into an initial probabilistic forecast over climatological quantile bins derived from historical observations or from the model's own hindcasts. Next, two lightweight machine learning models—Persistence++ and Debias++—are applied in parallel to produce complementary corrections of the initial forecast. Debias++ perturbs the forecast using site-, date-, and quantile-specific corrections learned by minimizing probabilistic forecasting error over adaptively-selected training periods. Meanwhile, Persistence++ blends the forecast with lagged observations and climatology to account for recent weather trends and day-of-year effects. Persistence++ and Debias++ are further described in Supplement A.1 and A.2. The two corrections are projected onto the space of valid distributions. Finally, they are averaged together so that the final PBC prediction takes into account both sources of information.

The PBC framework is lightweight computationally. Training and inference combined for all three forecast variables across both time horizons can be done using 12 CPUs in under 5 hours.

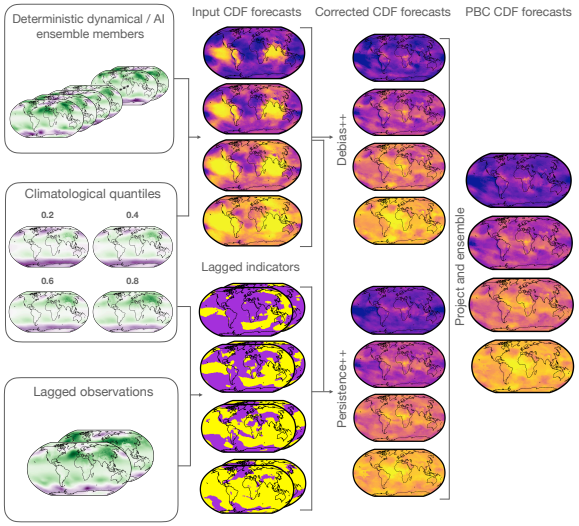

Figure 1: Probabilistic bias correction pipeline. In the forecast maps, purple corresponds to lower probabilities while yellow corresponds to higher probabilities.

The raw CDF predictions from PBC may violate the monotonicity constraint that cumulative probabilities must be non-decreasing across quintile thresholds. To ensure valid probabilistic forecasts, we apply isotonic regression as a post-processing step. For each grid point and target date, we first clip raw predictions to the valid probability range of zero to one, then apply isotonic regression to find the monotonically non-decreasing sequence that is closest in mean squared error to the input sequence, guaranteeing a valid CDF while minimally perturbing the original predictions.

As a post-processing system, PBC can be applied indiscriminately on the outputs of any forecasting model, dynamical or AI-based. We first apply PBC to the dynamical model from the European Centre for Medium-Range Weather Forecasts (ECMWF). Hereafter, we refer to the resulting corrected model as PBC-ECMWF. We also apply PBC to the Post-processing Ensembles with Transformers (PoET) forecasting model (Bouallègue et al., 2024), and refer to the resulting corrected model as PBC-PoET.

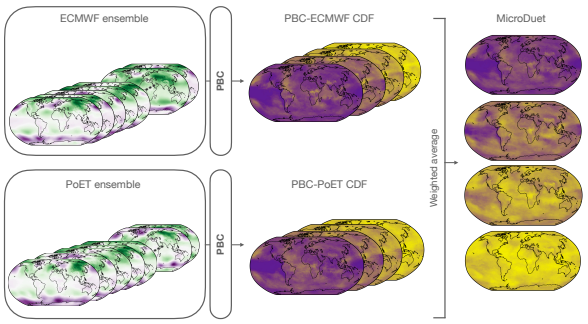

Figure 2: Duet model pipeline. In the forecast maps, purple corresponds to lower probabilities while yellow corresponds to higher probabilities.

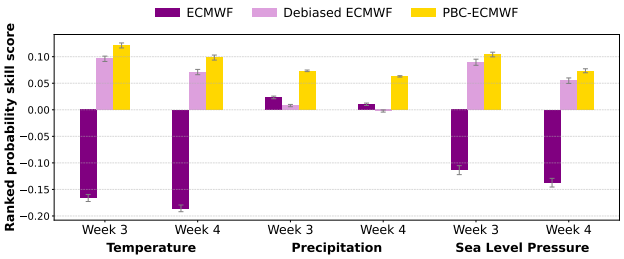

Figure 3: **Average forecast skill of the leading dynamical model (ECMWF) and its probabilistic bias correction (PBC).** Across the globe and the years 2016–2024, PBC boosts ECMWF skill more effectively than operational debiasing protocols for all variables and lead times.

2.2 Duet Model

The Duet model (Figure 2) forms a variable-specific weighted average of the PBC-ECMWF and PBC-PoET forecast outputs designed to leverage the complementary strengths of each. MicroDuet gives equal weight to PBC-ECMWF and PBC-PoET for temperature, full weight to PBC-ECMWF for precipitation, and full weight to PBC-PoET for mean sea level pressure. This variable-specific approach recognizes that ECMWF and PoET employ different forecasting schemes, leading to partially independent forecast errors. For temperature, averaging the two PBC-corrected predictions achieves error cancellation and improved reliability. For precipitation and sea level pressure, ensembling via simple average provides no additional benefit, and we therefore use the stronger single-model forecast.

3 Results

For concreteness, we adopt the following conventions of the ECMWF AI Weather Quest subseasonal forecasting competition (Loegel et al., 2025). First, we focus on forecasting average 2-m temperature (over land), total precipitation (over land), and average mean sea level pressure at a resolution of 1.5° latitude-longitude grid with 5 quintile bins. Secondly, we evaluate subseasonal forecasts for week 3 (days 19–25) and week 4 (days 26–32) using ranked probability skill score (RPSS) (Epstein, 1969), a standard measure of forecast improvement over a climatological baseline. Additional details on the methods, forecasting targets, and evaluation are described in Supplemental Materials.

3.1 Correcting Dynamical Forecasts with PBC

Figure 3 summarizes the significant skill gains obtained when PBC is applied to the industry-leading dynamical ensemble from ECMWF. For each target variable and lead time, it reports average global RPSS over the years 2016–2024. Remarkably, PBC boosts the subseasonal RPSS of ECMWF by 70–80% for precipitation and over 200% for temperature and mean sea level pressure, transforming low- and no-skill forecasts into skillful predictions that consistently outperform climatology. Analogous gains are observed when skill is stratified by season (Figure 7).

To dampen the systemic effects of model bias, forecasting centers routinely debias their operational forecasts by comparing each deterministic prediction to quantiles derived from model hindcasts rather than observations (ECMWF, 2026). Given the same deterministic predictions as input, PBC consistently outperforms this operational debiasing protocol, boosting debiased ECMWF precipitation RPSS by 89–103%, temperature RPSS by 21–28%, and pressure RPSS by 14–25% (Figure 3). The PBC gains over debiased ECMWF are also broadly distributed with 98% of grid cells showing skill gains for precipitation, 91–92% for temperature, and 89–91% for sea level pressure (Figure 6). These results suggest that PBC can serve as a more skillful drop-in replacement for existing operational debiasing protocols.

In fact, PBC can boost the skill of its dynamical input beyond the levels of existing AI-based subseasonal models. For example, Chen et al. (2024) recently introduced FuXi-S2S, a deep learning model that outperforms ECMWF's dynamical ensemble at subseasonal lead times. However, across the globe and the test years 2017–2021, PBC-ECMWF improves upon FuXi-S2S RPSS by over 200% for every variable and lead time, with 99–100% of grid cells showing skill gains for temperature, 96–97% for precipitation, and 93% for mean sea level pressure (Figure 8).

3.2 Probabilistic bias correction of AI models

PBC can also be used to improve the skill of AI models. We apply PBC to a hybrid (AI + dynamical) forecasting system, PoET. PoET takes as input a dynamical ensemble (in this case, ECMWF) and outputs a new ensemble with all members simultaneously corrected by a transformer neural network (Bouallègue et al., 2024). PoET

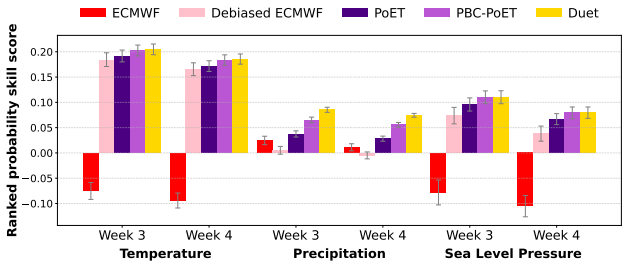

Figure 4: **Average forecast skill of the PoET AI model and its probabilistic bias correction (PBC) in 2024.** Across the globe, PBC consistently boosts PoET skill for each weather variable and lead time. Duet yields further skill gains by ensembling PBC-PoET and PBC-ECMWF. The error bars display 95% bootstrap confidence intervals.

serves as challenging input for PBC as its ensemble has already been corrected by AI and is substantially more skillful than ECMWF itself. Nevertheless, PBC enhances PoET skill for every variable and lead time with RPSS gains of 41–49% for precipitation, 13–19% for mean sea level pressure, and 5–6% for temperature (Figure 4).

In 2025, ECMWF launched the AI Weather Quest, an international competition to benchmark and improve the state of operational subseasonal forecasting. Each week, contestants submit real-time probabilistic forecasts for weather around the globe, and prizes are awarded at the end of each season for the highest week 3 and week 4 skill. To benchmark the real-time operational utility of PBC, we the Duet model into the competition under the name MicroDuet. In the inaugural September-October-November season, MicroDuet placed first for all weather variables and lead times, outperforming the debiased dynamical forecasts of six government agencies (ECMWF, NOAA, ECCC, HMCR, KMA, and JMA), a multi-model ensemble of these six dynamical forecasts, three versions of ECMWF's AI forecasting system, and the forecasting systems of 34 teams worldwide.

3.3 Extreme event forecasting

An important application of subseasonal forecasting is the early identification of extreme weather to support preparedness for hazards such as floods, heatwaves, droughts, storms, and cold snaps. Figure 5 summarizes the significant gains in week 4 extreme forecasting skill when PBC is applied to the industry-leading dynamical ensemble from ECMWF. Here, we define extreme events as those verifying in the lowest or highest quintiles and measure skill using the Brier skill score (BSS) (Brier, 1950), a standard measure of improvement over a climatological baseline for a binary event.

Remarkably, PBC boosts the week 4 BSS of ECMWF by 27–91% for precipitation extremes, 150–772% for temperature extremes, and 262–656% for mean sea level pressure extremes, again transforming low- and no-skill forecasts into skillful predictions that consistently outperform climatology (Figure 5a). Given the same deterministic predictions as input, PBC also consistently outperforms operationally debiased ECMWF, with gains of 62–76% for precipitation extremes, 31–35% for temperature extremes, and 26–49% for pressure extremes (Figure 5a). The PBC improvements over debiased ECMWF are also broadly distributed with 96–97% of grid cells showing skill gains for precipitation, 87–92% for temperature, and 87–94% for sea level pressure (Figure 9).

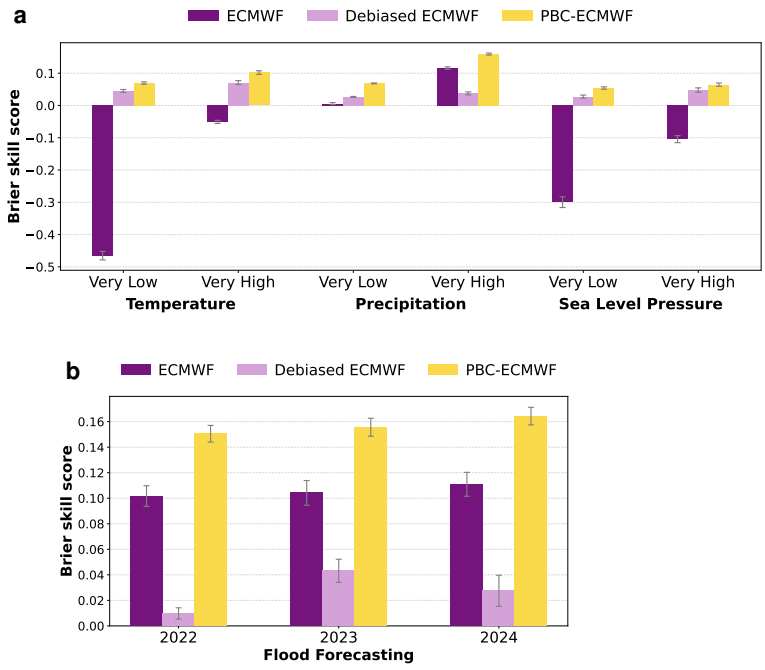

Figure 5: **Forecasting extreme weather with the leading dynamical model (ECMWF) and its probabilistic bias correction (PBC) in week 4.** (**a**) Across the globe and the years 2016–2024, PBC boosts ECMWF extreme weather forecasting skill more effectively than operational debiasing protocols for every variable and extreme. (**b**) Across flood events catalogued by the Global Disaster Awareness and Coordination System, operational debiasing severely degrades flood forecasting skill (top precipitation quintile Brier skill score in the $20° \times 20°$ bounding box centered on the flood), while PBC consistently enhances it.

To evaluate whether these gains translate into improved hazard prediction, Figure 5b displays average flood forecasting skill across all flood events catalogued by the Global Disaster Alert and Coordination System (GDACS) (De Groeve et al., 2006) in the years 2022–2024. While operational debiasing severely degrades the flood forecasting skill of the raw ECMWF ensemble, PBC consistently

enhances skill, boosting BSS by 72–94% over debiased ECMWF and 32% over ECMWF each year. Comparable improvements are also observed in week 3 (Figure 10). These results demonstrate that PBC not only improves subseasonal forecast quality but also enhances early warning capability for high-impact extreme events.

4 Discussion

Our results indicate that probabilistic post-processing with machine learning can substantially boost the quality and utility of subseasonal forecasts. The probabilistic bias correction (PBC) framework achieves these goals by satisfying four key properties. First, across the globe, PBC consistently improves the world's leading dynamical model from ECMWF, yielding a new state of the art for hybrid (AI + dynamical) subseasonal forecasting. Second, PBC consistently improves leading subseasonal AI models, establishing a new state of the art for fully data-driven subseasonal forecasting. Third, PBC consistently improves the early detection of extreme weather events. Fourth, PBC forecasts are suitable for operational deployment and, in real-time competition, consistently outperform the best alternative dynamical, AI, and hybrid forecasting systems from around the world.

PBC targets predictable errors rather than attempting to relearn full atmospheric dynamics. Hence, it should be viewed as an effective post-processing tool rather than a replacement for model improvement. Fortunately, the PBC framework is adaptive: as dynamical and AI models are upgraded with improved scientific understanding, process representations, or training data, PBC can be retrained to ingest and correct these higher-quality inputs. PBC's low cost provides an efficient, low-overhead strategy for meteorological centers to enhance their existing forecast products.

Overall, our results open the door to new advances in extended-range forecasting. With subseasonal lead times, PBC already has the potential to benefit a myriad of downstream applications, including agriculture, energy production, healthcare, and disaster response, and the same approach can be applied to longer, e.g., seasonal outlooks. Through the accompanying release of our open-source code, we aim to provide a standardized tool for the community to bridge the gap between raw model output and actionable intelligence at subseasonal horizons and beyond.

References

Bouallègue, Z. B., Weyn, J. A., Clare, M. C. A., Dramsch, J., Dueben, P., and Chantry, M. (2024). Improving medium-range ensemble weather forecasts with hierarchical ensemble transformers. *Artificial Intelligence for the Earth Systems*, 3(1):e230027.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Chen, L., Zhong, X., Li, H., Wu, J., Lu, B., Chen, D., Xie, S.-P., Wu, L., Chao, Q., Lin, C., et al. (2024). A machine learning model that outperforms conventional global subseasonal forecast models. *Nature Communications*, 15(1):6425.

De Groeve, T., Peter, T., Annunziato, A., and Vernaccini, L. (2006). Global disaster alert and coordination system. In *Proceedings of the 3rd International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, Newark, NJ, USA. ISCRAM.

ECMWF (2026). SUBS-M-climate, the subseasonal model climate. https://web.archive.org/web/20260222173505/https://confluence.ecmwf.int/display/FUG/Section+5.3.2+SUBS-M-climate%2C+the+sub-seasonal+model+climate. Accessed: 2026-02-22.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology and Climatology*, 8(6):985–987.

Loegel, O., Talib, J., Vitart, F., Hoffmann, J., and Chantry, M. (2025). The ai weather quest: an international competition for sub-seasonal forecasting with ai. *Machine Learning: Earth*, 1(1):010701.

Lorenz, E. N. (1969). The predictability of a flow which possesses many scales of motion. *Tellus*, 21(3):289–307.

Merryfield, W. J., Baehr, J., Batté, L., Becker, E. J., Butler, A. H., Coelho, C. A., Danabasoglu, G., Dirmeyer, P. A., Doblas-Reyes, F. J., Domeisen, D. I., et al. (2020). Current and emerging developments in subseasonal to decadal prediction. *Bulletin of the American Meteorological Society*, 101(6):E869–E896.

Mouatadid, S., Orenstein, P., Flaspohler, G., Cohen, J., Oprescu, M., Fraenkel, E., and Mackey, L. (2023). Adaptive bias correction for improved subseasonal forecasting. *Nature Communications*, 14(1).

Nathaniel, J., Qu, Y., Nguyen, T., Yu, S., Busecke, J., Grover, A., and Gentine, P. (2024). Chaosbench: A multi-channel, physics-based benchmark for subseasonal-to-seasonal climate prediction. *Advances in Neural Information Processing Systems*, 37:43715–43729.

Richter, J. H. and Joseph, E. (2025). Scientists must join forces to solve forecasting's predictability desert. *Eos*, 106.

Bridging the "Predictability Desert": A Probabilistic Bias Correction Framework for AI and Dynamical Subseasonal Forecasts Supplementary Materials

A Model details

A.1 Persistence++

Persistence++ is inspired by classical persistence forecasting by using lagged features to account for recent weather trends up to the forecast issuance date. For each grid point and quintile threshold, it regresses binary targets (whether the observed value falls at or below the threshold) onto three types of features: 1) binary indicators showing whether past observations fell at or below the threshold, using two specific lags: one at the ground truth availability lag and another accounting for the forecast horizon length, 2) the raw ensemble forecast from the underlying model (dynamical or AI) to provide physical guidance, and 3) a rolling climatology to provide historical expectations.

For each quintile threshold, grid point, and target date, a Persistence++ model is trained using all available historical forecasts and observations observable as of the forecast issuance date.

A.2 Debias++

Debias++ is a three-step approach to ensemble forecast correction. The method adaptively selects optimal training windows to ensemble and debias input forecasts based on recent historical performance.

For each target date and quintile threshold, Debias++ first adaptively selects a window of observations around the target day of year and a range of forecast issuance dates and lead times for ensembling. This selection is based on which configuration achieved the smallest latitude-weighted mean squared error over the preceding 3 years.

The method considers multiple configurations that vary in: (i) the span of calendar days included on either side of the target day of year, (ii) the number of forecast issuance dates to average over, and (iii) which lead times to include in the ensemble average. By varying these parameters, Debias++ can adapt to heterogeneity in forecast errors over time.

Once the optimal configuration is selected, Debias++ forms an ensemble mean forecast by averaging the raw CDF predictions over the selected range of issuance dates and lead times. Finally, for each grid point, Debias++ bias-corrects the ensemble forecast by subtracting the historical error over the last 20 years, i.e., subtracting the mean ensemble forecast over the selected window of observations from the mean observed probability . This bias correction step removes systematic over- or under-prediction that may be present in the raw ensemble forecasts.

Unlike standard debiasing strategies which employ static ensembling and bias correction parameters, Debias++ adapts to forecast error heterogeneity by learning to vary the amount of temporal ensembling and the size of the observation window over time, allowing it to provide location-, season-, and quintile-specific corrections.

B Additional results

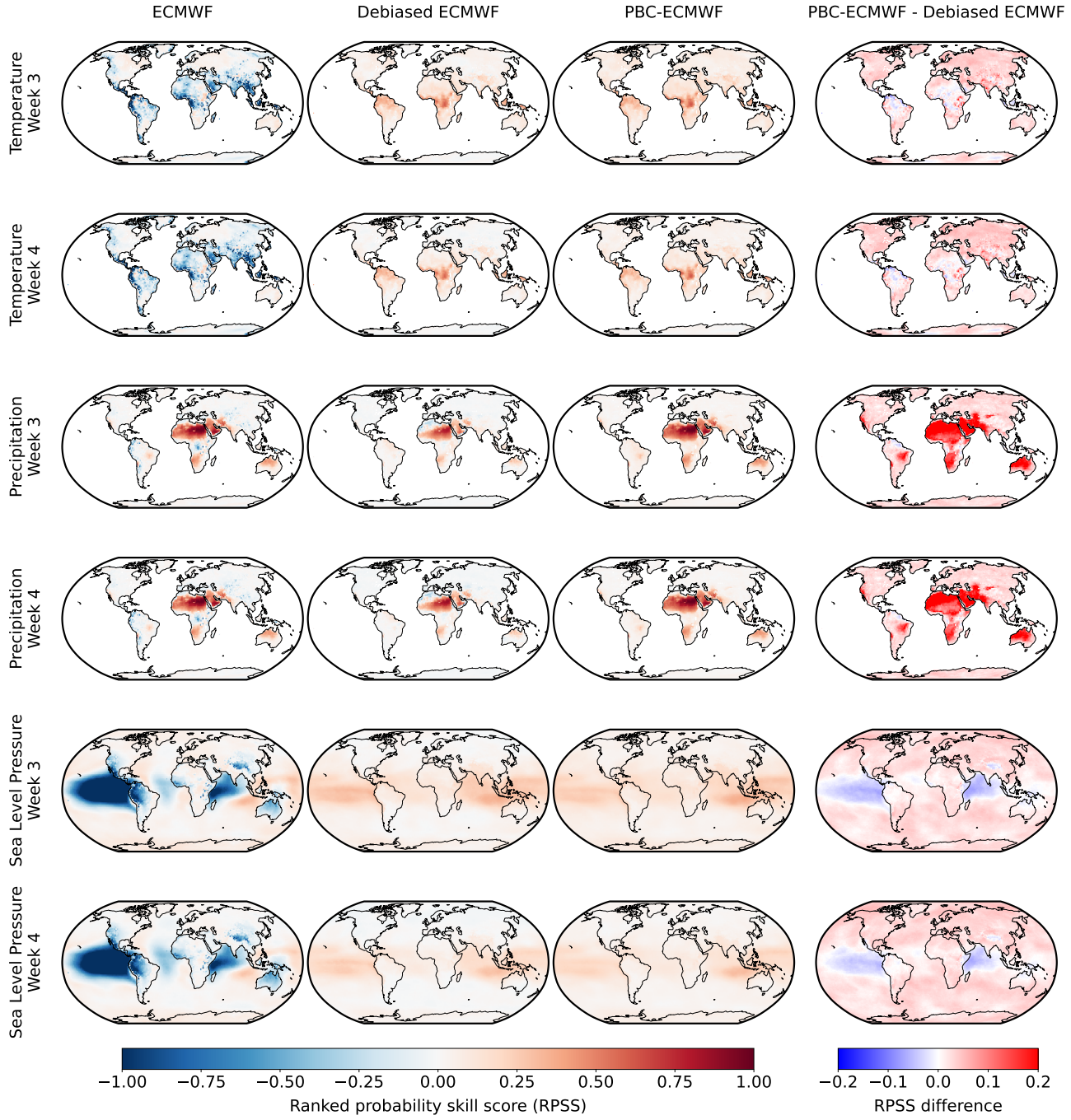

Figure 6: **Spatial distribution of skill of the leading dynamical model (ECMWF) and its probabilistic bias correction (PBC).** Across the globe and the years 2016–2024, PBC consistently boosts the skill of ECMWF more effectively than operational debiasing protocols. These improvements are broadly distributed with 98% of grid cells showing skill gains over debiased ECMWF for precipitation, 91–92% for temperature, and 89–91% for sea level pressure.

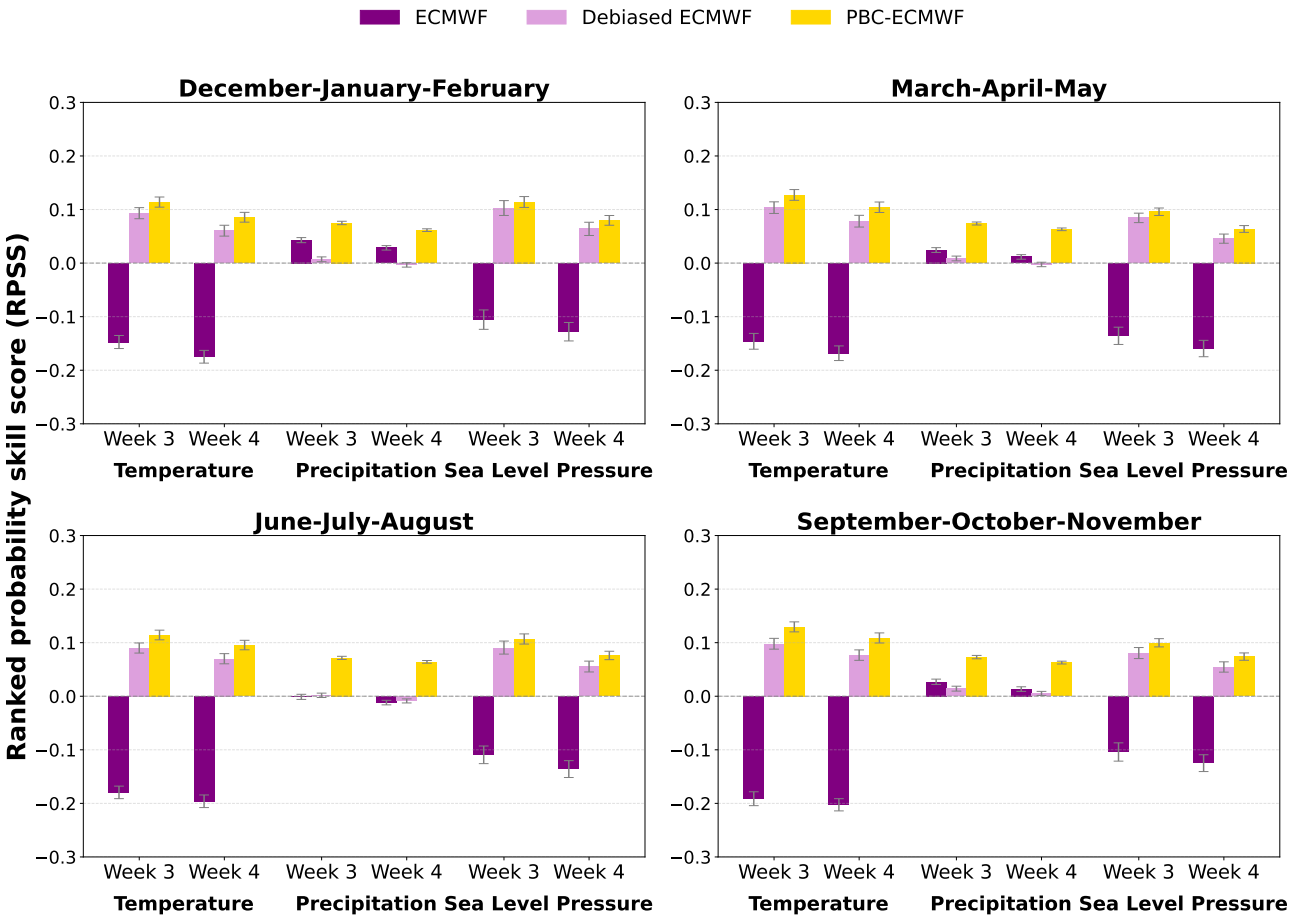

Figure 7: **Seasonally-averaged forecast skill of the leading dynamical model (ECMWF) and its probabilistic bias correction (PBC).** Across the globe and the years 2016–2024, PBC boosts the skill of ECMWF and operationally debiased ECMWF in every season.

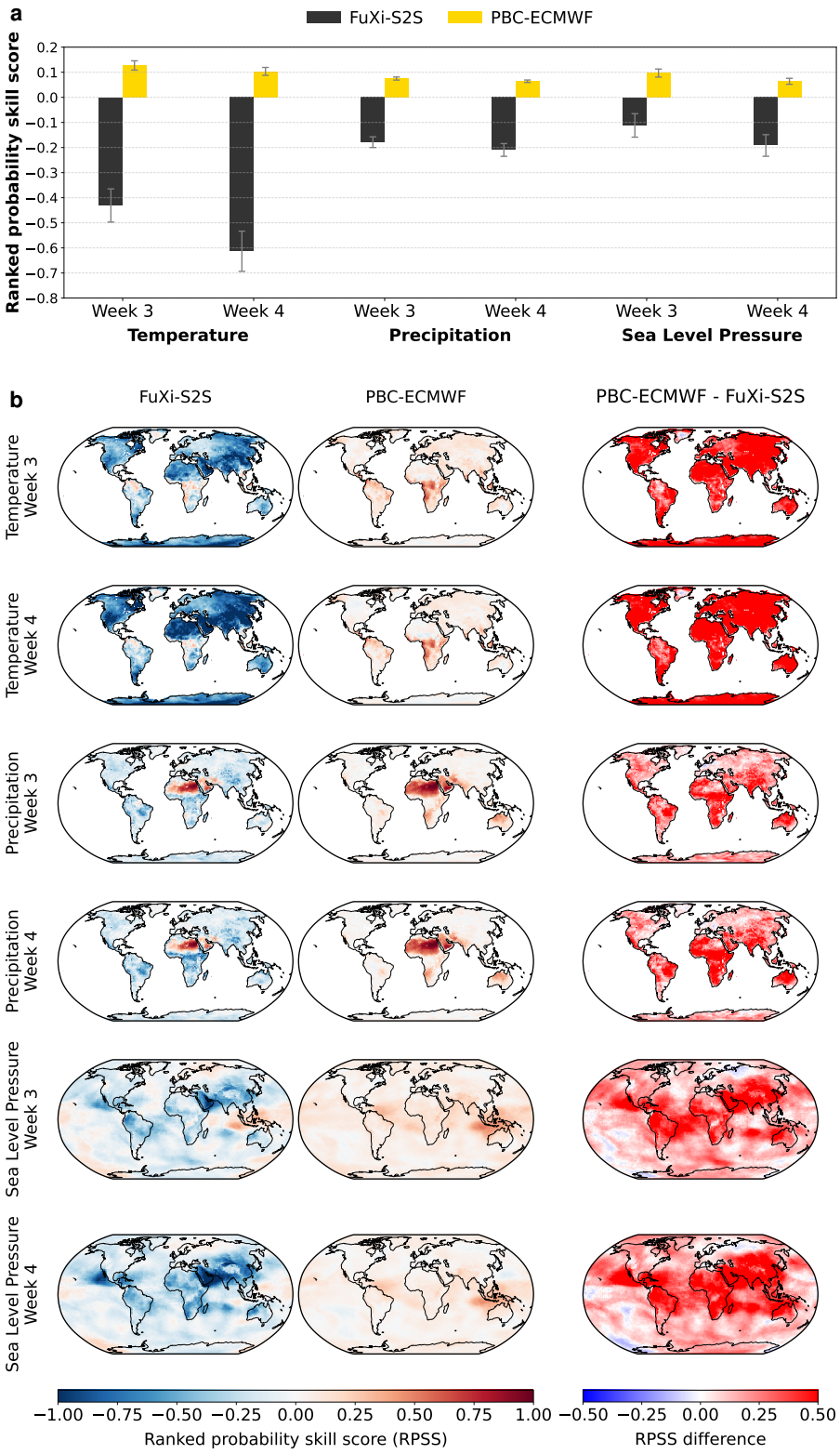

Figure 8: **Average forecast skill of FuXi-S2S and PBC-ECMWF.** Across the globe and the years 2017–2021, PBC-ECMWF **(a)** improves upon the subseasonal skill of the FuXi-S2S AI model by over 200% for every variable and lead time, with **(b)** 99–100% of grid cells showing skill gains for temperature, 96–97% for precipitation, and 93% for mean sea level pressure. The error bars display 95% bootstrap confidence intervals.

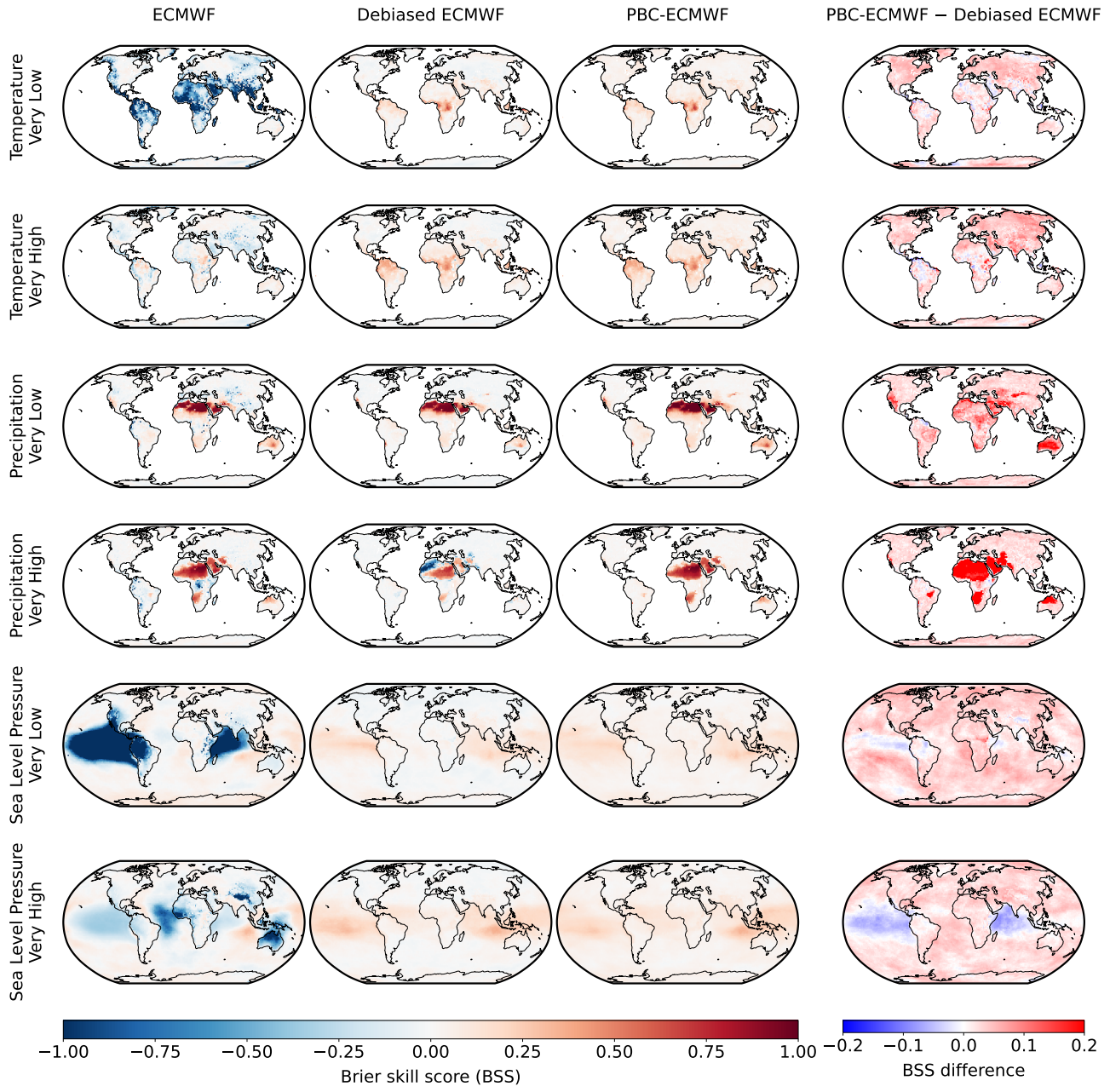

Figure 9: **Forecasting extreme weather with the leading dynamical model (ECMWF) and its probabilistic bias correction (PBC) in week 4.** Taking the same dynamical forecasts as input, PBC also improves over operational debiasing protocols, with 96–97% of grid cells showing skill gains for precipitation, 87–92% for temperature, and 87–94% for sea level pressure.

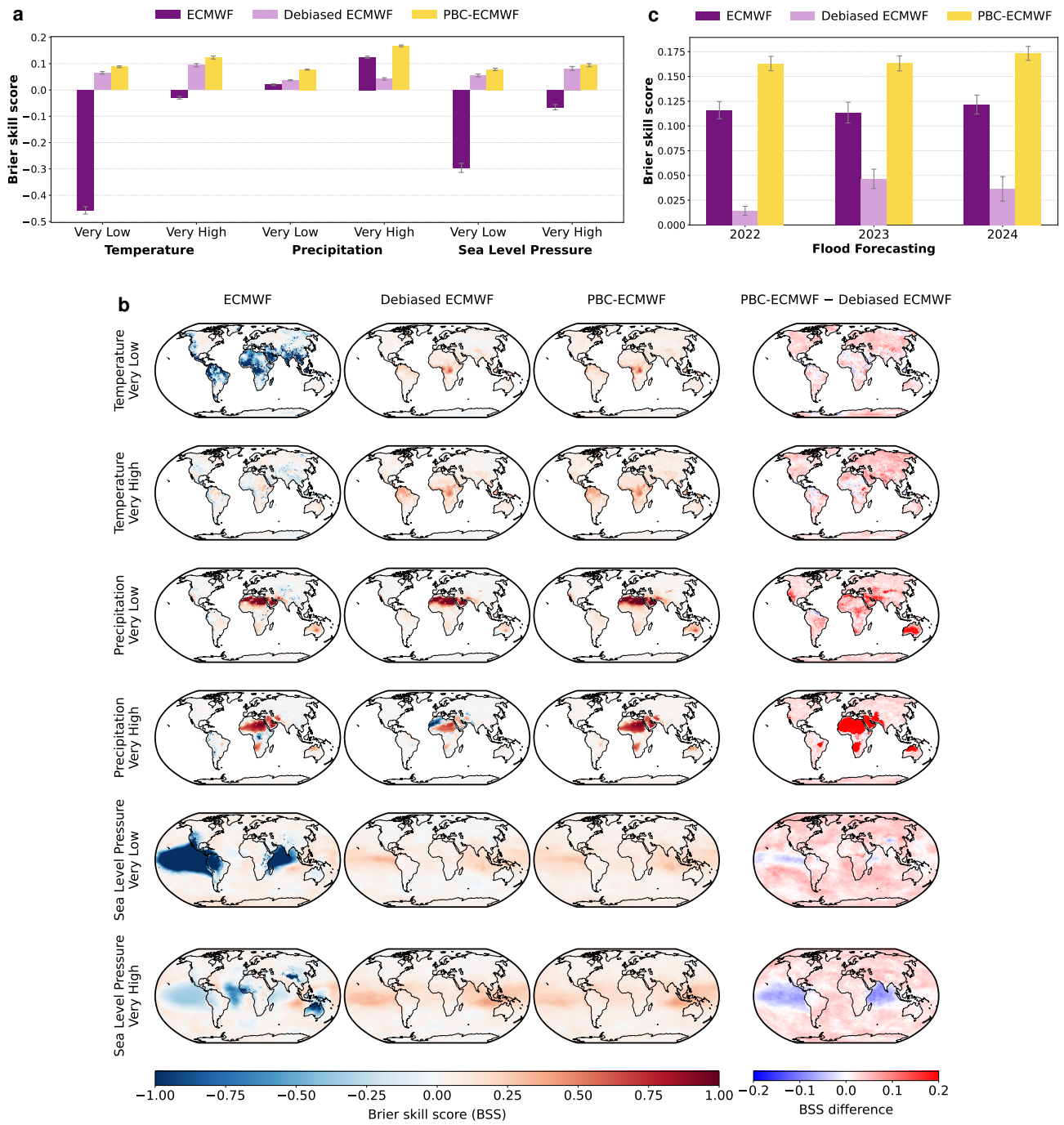

Figure 10: **Forecasting extreme weather with the leading dynamical model (ECMWF) and its probabilistic bias correction (PBC) in week 3.** (**a**) Across the globe and the years 2016–2024, PBC boosts ECMWF forecast skill by 124–619% for temperature extremes, 26–74% for precipitation extremes, and 169–478% for mean sea level pressure extremes. Taking the same dynamical forecasts as input, PBC also improves over operational debiasing protocols, (**a**) boosting the extreme forecast skill of debiased ECMWF by 52–75% for precipitation, 24–26% for temperature, and 14–29% for pressure and (**b**) gaining skill in 87–97% of sites worldwide. (**c**) Across flood events catalogued by the Global Disaster Awareness and Coordination System, PBC improves extreme precipitation skill by 72–91% over debiased ECMWF and 29% over ECMWF each year. The error bars display 95% bootstrap confidence intervals.