Improving Subseasonal to Seasonal Global 2-m Temperature and Precipitation Forecasts Using a Deep Learning–Based Postprocessing Approach

CHALACHEW KINDIE MENGIST[a,b] AND KYONG-HWAN SEO[a,b,c]

[a] *Institute for Future Earth, Pusan National University, Busan, South Korea*
[b] *BK21 School of Earth and Environmental Systems, Pusan National University, Busan, South Korea*
[c] *Research Center for Climate Sciences, Department of Atmospheric Sciences, Pusan National University, Busan, South Korea*

(Manuscript received 4 August 2025, in final form 31 December 2025, accepted 19 January 2026)

ABSTRACT: Numerical weather prediction models continue to exhibit flaws in forecasting subseasonal to seasonal (S2S) 2-m temperature (T2M) and total precipitation (TP) beyond 2 weeks. In this study, a deep learning–based postprocessing model for the S2S probabilistic forecast of global T2M and TP was developed using a residual U-Net (ResU-Net) algorithm. The model was developed based on deterministic input information from weekly reforecasts of the European Centre for Medium-Range Weather Forecasts (ECMWF); these aggregates included five upper-air atmospheric variables at three pressure levels (250, 500, and 850 hPa) and six surface variables. ResU-Net produced global probabilistic tercile forecasts of T2M and TP for weeks 1–6, including biweekly aggregates of weeks 3–4 and 5–6. Compared with the ECMWF S2S forecasts, the ResU-Net probabilistic forecast demonstrated superior probabilistic forecasts for up to 6 weeks. The most important predictors to forecast T2M in lead time weeks 1 and 2 are above mean sea level (MSL), 850-hPa geopotential height (Z850), 850-hPa temperature (T850), and T2M, while weeks 4–6 are sea surface temperature (SST), 200-hPa geopotential height (Z200), and 200-hPa temperature (T200). For TP forecast, TP and total column water are important predictors during weeks 1–3 and TP and SST are important predictors for weeks 4–6. The ocean–atmosphere coupling and large-scale teleconnections influence longer lead week forecasts of T2M and TP. Therefore, combining numerical weather prediction models and deep learning show promising S2S probabilistic forecasts.

KEYWORDS: Surface temperature; Precipitation; Forecast verification/skill; Probability forecasts/models/distribution; Seasonal forecasting; Deep learning

1. Introduction

Subseasonal to seasonal (S2S) forecasting provides weather phenomena between 2 weeks and 3 months. Skillful S2S prediction assists decision-making sectors in public health, agriculture, energy supply, and water resource management in formulating advance plans for addressing extreme weather conditions, such as heat waves, droughts, floods, and cold spells (Domeisen et al. 2022; Pegion et al. 2019; White et al. 2017, 2022). Unlike short- to medium-range weather forecasting systems, physics-based S2S numerical weather prediction models have lower (or better) forecasting skills than climatological (or persistence) forecasting (Robertson et al. 2020). The S2S forecast is challenging due to its dependence on both local weather and global climate variables (Merryfield et al. 2020), the chaotic nature of the atmosphere, and the insufficient input information from land beyond 1 month.

Postprocessing machine learning (ML) algorithms, such as random forest (Taillardat et al. 2016), gradient boosting (Messner et al. 2017), neural networks (Rasp and Lerch 2018), and adaptive bias correction (Mouatadid et al. 2023), have been used to improve the accuracy of short- to medium-range weather forecasting (Vannitsem et al. 2021) and climate predictions (Ham et al. 2019). Several postprocessing applications have utilized convolutional neural network (CNN)-based architectures to extract and learn spatial error structures of short- to medium-range forecasts (Gröenquist et al. 2021; Veldkamp et al. 2021; Chapman et al. 2022; Hu et al. 2023). Major technology companies developed ML models to improve the short- to medium-range (0–10 days) forecasting accuracy. NVIDIA has developed the transformer-based Fourier Forecasting Neural Network (FourCastNet) (Pathak et al. 2022; Bonev et al. 2023) with spherical Fourier neural operators (SFNOs), Google has created the Deepmind GraphCast (Lam et al. 2023) with graph neural network (GNN), Huawei has released the Swin-Transformer based Pangu-Weather model (Bi et al. 2023), and the European Centre for Medium-Range Weather Forecasts (ECMWF) has developed GNN-based Artificial Intelligence/Integrated Forecasting System (AIFS) (Lang et al. 2024). Recently, Google introduced NeuralGCM (Kochkov et al. 2024), a novel technique that combines partial differential equations and machine learning to generate ensemble weather predictions for 1–15-day forecasts.

Although minimal research has been conducted on S2S postprocessing approaches (Robertson et al. 2020), some researchers have developed ML and deep learning techniques to enhance the forecasting skills of S2S predictions. Scheuerer et al. (2020) developed a CNN-based architecture to forecast S2S precipitation over California. Machine learning [extreme gradient boosting (XGBoost) and multitask Lasso] and deep learning {encoder [long short-term memory (LSTM)]–decoder [feed-forward neural network (FNN)]} algorithms previously outperformed climatological baselines in forecasting S2S 2-m temperatures (T2M) and total precipitation (TP) over the U.S. mainland for weeks 3 and 4 (He et al. 2021). The S2S forecast skill of the ECMWF for weeks 1–4 summer precipitation over

Corresponding author: Kyong-Hwan Seo, khseo@pusan.ac.kr

DOI: 10.1175/MWR-D-25-0179.1

southern China is improved using U-Net-based deep learning (Lyu et al. 2023). FuXi-S2S (Chen et al. 2024) developed a deep learning–based model to forecast global atmospheric variables, the Madden–Julian oscillation (MJO), and precipitation from reanalysis data. In 2021, the World Meteorological Organization (WMO) held an artificial intelligence (AI) challenge to improve global S2S forecasts of T2M and TP (Vitart et al. 2022).

The 2021 WMO AI challenge mainly used ML algorithms utilizing mostly gridcell-specific approaches, including ensemble model output statistics, logistic regression, and random forest classification (Vitart et al. 2022). In the WMO AI challenge, T2M and TP serve as both predictors and predictands for biweekly forecasts at lead times of weeks 3–4 and 5–6. Horat and Lerch (2024) reported S2S probabilistic skill forecasts using a U-Net-based postprocessing method for weeks 3–4 and 5–6 of T2M and TP with four and five predictors, respectively. However, neither Vitart et al. (2022) nor Horat and Lerch (2024) reported weekly probabilistic forecast skills. Furthermore, previous studies have not reported the contributions of other predictors from upper-air atmospheric and surface variables, particularly the slowly varying components of the climate system. Therefore, this study incorporated multiple predictor variables to train with the ResU-Net algorithm, with the goals of enhancing the probabilistic forecast accuracy for T2M and TP for lead times of weeks 1–6, weeks 3–4, and weeks 5–6; analyzing the importance of each predictor across lead times; and evaluating the seasonal forecast skill of T2M and TP. The manuscript is organized as follows: Section 2 provides details about the S2S AI challenge data used in this study and ResU-Net. Section 3 describes the deep learning forecast results for the study period. Finally, the conclusions of the study are provided in section 4.

2. Data and methodology

a. Data

Similar to the WMO S2S AI challenge, the ECMWF (Vitart et al. 2017) forecast and reforecast and the NOAA Climate Prediction Center (CPC) observation data are used in this study. The ECMWF weekly ensemble reforecasts with 11 members from the years 2000 to 2019 are used for training, and 51-member forecasts of the year 2020 serve as testing. Owing to the ensemble size difference between reforecasts and forecasts, the ensemble mean for both the reforecast and forecast datasets is calculated, and the postprocessing technique is applied.

The ECMWF reforecasts and forecasts had a 1.5° horizontal resolution and weekly initialization for 46 daily temporal resolutions. The weekly aggregates of the forecasts are calculated by averaging (summing) the daily values for a given variable for weeks 1–6, 3–4, and 5–6, as described by Vitart et al. (2022). For example, we derived weekly forecasts of T2M (TP) for weeks 1–6 from 7-day rolling window averages (sum) with lead times of 0, 7, 14, 21, 28, and 35 days, respectively, whereas we derived the weeks 3–4 and 5–6 T2M (TP) forecasts from 14-day rolling window averages (sum) with lead times of 14 and 28 days, respectively.

The weekly probability forecasts for "below normal," "near normal," and "above normal" values for the target variables, T2M and TP, were computed based on terciles of observation data from the NOAA CPC verification during the training year (2000–19) (Vitart et al. 2022; Horat and Lerch 2024). For the tercile probability forecasts, the distance from the ensemble mean of the target variable forecast to the respective tercile category edges was calculated. These distances to the upper and lower boundaries of the middle tercile are two predictors (Horat and Lerch 2024), which are more informative than the ensemble mean. Hereafter, T2M (TP) upper and T2M (TP) lower denote the distances from the ensemble-mean forecast to the upper and lower tercile boundaries, respectively. For nontarget variables, anomaly values for the ensemble mean are computed for global surface variables of above mean sea level (MSL) pressure, total column water vapor (TCW), total cloud cover (TCC), top net thermal radiation (TTR), and sea surface temperature (SST); the upper-air atmospheric variables at 250, 500, and 850 hPa include geopotential height, temperature, zonal component of wind, meridional component of wind, and specific humidity as predictors, and probability forecasts for T2M and TP are the predictands. We used the same set of 22 input predictors for both training (2000–19) and test (2020) periods when predicting T2M and TP.

b. ResU-Net algorithm

In this study, we used ResU-Net (Fig. 1), a convolutional neural network–based deep learning algorithm that integrates the U-Net with residual network (ResNet) blocks. The U-Net used in this study has the same overall architecture as the ResU-Net but without the residual (skip) connections within each block. The inclusion of residual connections speeds up training and improves feature extraction compared to plain U-Net. The ResU-Net network has two main parts: (i) the encoder and (ii) decoder. Each ResNet block comprises two convolutional layers, followed by batch normalization and exponential linear unit (ELU) activation (Clevert et al. 2016). The ResNet block has local skip (shortcut) connection, which directly adds the input of the block to its output. The local skip has one convolution and batch normalization. This shortcut connection helps to mitigate the vanishing gradient problem and allows the network to learn residual mappings, thereby improving convergence and performance. The ResNet block is followed by a max pooling layer. The decoder of the network employs transposed convolutions to upsample the feature maps. This path also includes concatenation layers that merge features from the corresponding encoder layers. Each decoder stage is followed by ResNet block to further refine the features. The final layer consists of a 1×1 convolution with a softmax activation function (Bridle 1989), producing the segmentation map with class probabilities.

For the ResU-Net algorithm, hyperparameter tuning and training were performed using a 10-fold cross-validation method and 20 years of data (2000–19). Each cross-validation period covers 2 years, such as 2000–01, 2002–03, and so on. Nine folds were

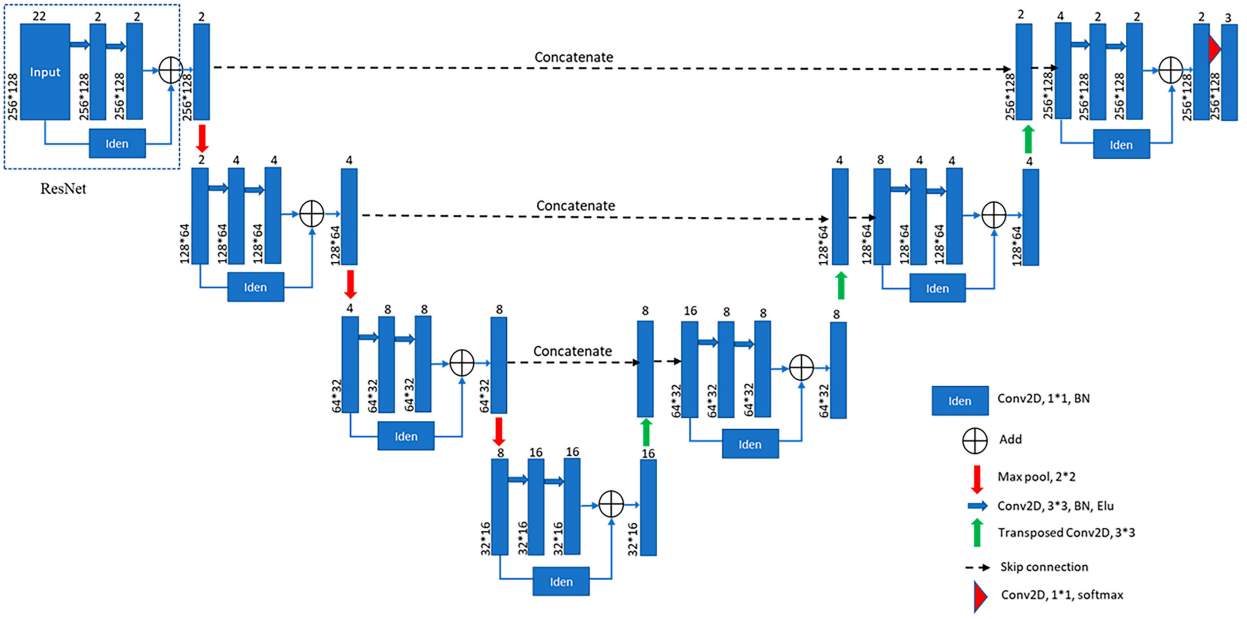

FIG. 1. Schematic diagram of ResU-Net.

used for model training, with the remaining utilized for validation. Preprocessing was conducted separately for the training and validation data in each split to avoid data leakage. The final forecast for the test year 2020 was obtained by averaging the predictions from the 10 models developed using 10-fold cross validation.

The model was trained using categorical cross-entropy loss averaged over all grid cells. The Adam optimizer (Kingma and Ba 2017) was used for training, and early stopping was applied based on validation loss to prevent overfitting.

c. *Forecast evaluation*

We evaluate the forecast skill of models (ECMWF and deep learning) with respect to the reference climatology using the ranked probability skill score (RPSS; Epstein 1969; Murphy 1971) for the test dataset (year 2020). Let F_m and O_m be the forecasts and observations with m being the number of probabilistic multicategories, respectively. In this study, the number of categories, m, are three (below normal, near normal, above normal).

The ranked probability score (RPS) is defined as

$$\mathrm{RPS} = \sum_{m=1}^{M} (F_m - O_m)^2. \tag{1}$$

A perfect forecast achieves an RPS of 0 (Wilks 2020). The RPSS relates the RPS of the forecast (RPS_f; ECMWF and deep learning) and climatology ($\mathrm{RPS}_{\mathrm{clm}}$). The climatology ($\mathrm{RPS}_{\mathrm{clm}}$) assigns equal probabilities of 1/3 to all three outcomes.

The RPSS is computed as

$$\mathrm{RPSS}(f) = 1 - \frac{\mathrm{RPS}_f}{\mathrm{RPS}_{\mathrm{clm}}}. \tag{2}$$

If RPSS is positive, the forecasting skills of ECMWF and deep learning are more skillful than climatology, and vice versa.

d. *Permutation importance*

To investigate the relative importance of each predictor, a permutation importance method described by Rasp and Lerch (2018) is utilized. A single predictor is randomly shuffled in test data at a time, while the remaining predictors are unchanged, and the most important predictor shows a decrease in RPSS of a permuted predictor compared with the mean of RPSS of unpermuted predictors. The feature importance of a predictor p (FI_p) is calculated as

$$\mathrm{FI}_p = \frac{1}{N}\sum_{n=1}^{N}[\mathrm{RPSS}_{\mathrm{base}} - \mathrm{RPSS}(n)_p], \tag{3}$$

where $\mathrm{RPSS}_{\mathrm{base}}$ and RPSS_p are the RPSS of the unpermuted and permuted version of the pth predictor, respectively, as defined by Rasp and Lerch (2018). A larger FI_p indicates greater importance of the pth predictor. To ensure robustness, each predictor is permuted 100 times.

3. Results

a. *Global forecast skill of T2M and TP*

Figure 2 shows an example of the three categories of below normal, near normal, and above normal of the T2M from observation, ECMWF, U-Net, and ResU-Net for the lead time of week 2, issued on 2 January 2020. For the TP prediction example, see the appendix (Fig. A1). Overall, the spatial structure across models is similar to the observations. However, U-Net and ResU-Net outputs are smoother and less

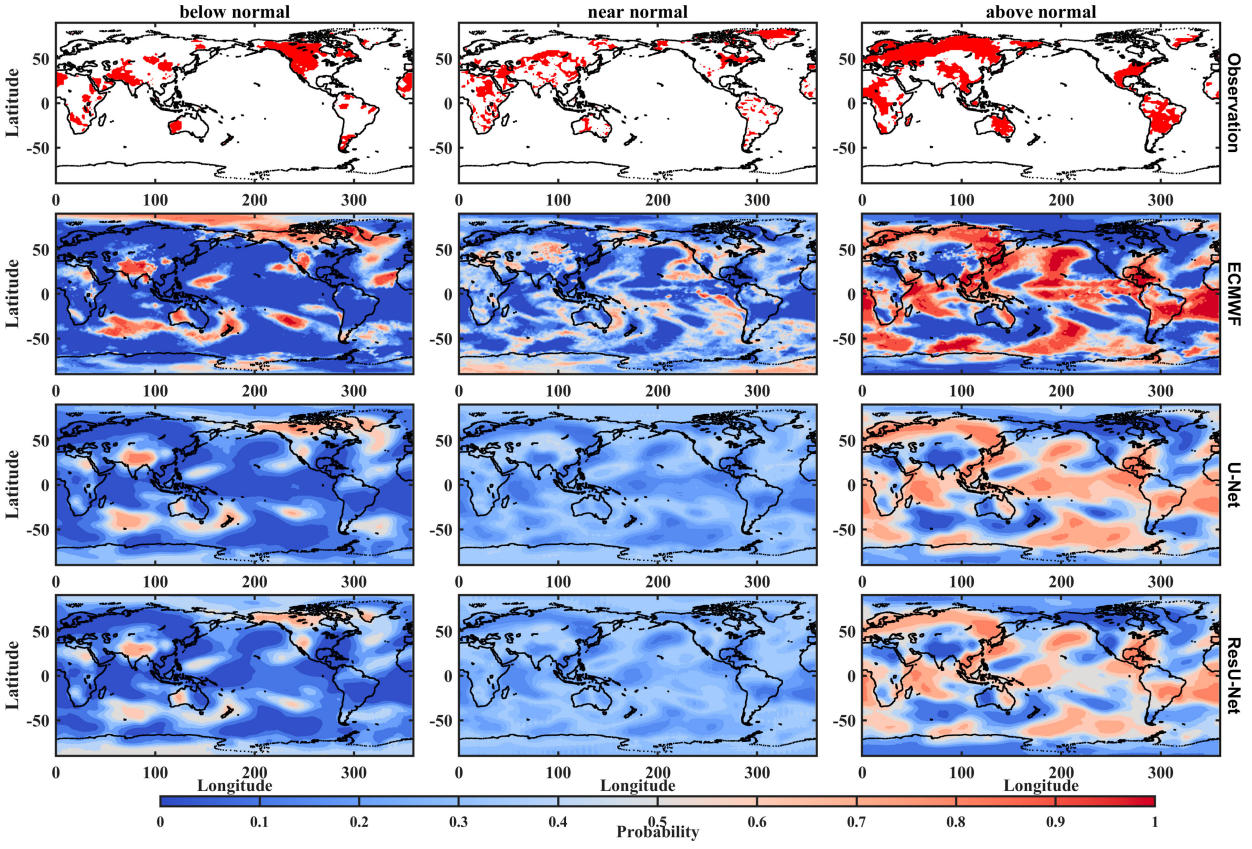

FIG. 2. Example of probabilistic predictions of T2M at a 2-week lead, issued on 2 Jan 2020. The tercile categories of (left) below-normal, (middle) near-normal, and (right) above-normal conditions, respectively. (first row) The observed category (red shading), followed by (second row) ECMWF, (third row) U-Net, and (fourth row) ResU-Net forecasts.

sharp than ECMWF. The deep learning probability outputs for near normal are close to climatology, which may be due to the distribution characteristics of the near-normal tercile and the small differences between the boundaries of the upper and lower tercile edges, consistent with the analysis presented by Horat and Lerch (2024).

The forecast skill of ECMWF and the deep learning model with respect to reference climatology was evaluated using the RPSS, as formulated by Epstein (1969) and Murphy (1971), for the independent test dataset (year 2020). Figure 3 presents the global RPSS for T2M and TP from the ResU-Net and ECMWF models for representative lead weeks 1, 3, and 5. The complete weekly results for all six lead weeks are shown in the appendix (Fig. A3). Positive RPSS values indicate that the probabilistic forecasts demonstrate superior skill relative to climatology, whereas negative values indicate lower predictive skill. The corresponding U-Net forecast skill for T2M and TP is provided in the appendix (Fig. A2).

Figure 3a depicts the first week (week 1) of the T2M forecast accuracy. The Northern Hemisphere (NH), particularly North America, Europe, and parts of Asia, show positive RPSS (red areas), indicating higher predictive performance. On the other hand, regions such as east of South America and East Africa show lower or negative RPSS (blue areas), indicating the challenges in forecasting T2M. In week 3 (Fig. 3b) and week 5 (Fig. 3c), overall accuracy decreases as lead time increases, though ResU-Net retains modest skill over several regions. ECMWF results (Figs. 3d–f) show consistently lower scores across all weeks, confirming the benefit of postprocessing.

The ResU-Net forecast skill for TP is also shown in Fig. 3, with analogous U-Net results in the appendix (Fig. A2). Forecast accuracy is highest in week 1 (Fig. 3g) and decreases steadily with lead time (Figs. 3h,i). Tropical regions were influenced by complex climate interactions, which resulted in a significant decrease in forecasting accuracy levels. While both T2M and TP forecast skill diminish with longer leads, ResU-Net maintains broader areas of positive RPSS compared with ECMWF.

Table 1 shows RPSS values of ECMWF, U-Net, and ResU-Net models for T2M and TP forecast skills in the NH, tropics, Southern Hemisphere (SH), and global. Deep learning methods outperform ECMWF forecasts in all weeks. U-Net performed slightly better in week 1, but ResU-Net showed more accuracy as lead time increased. ResU-Net was more stable over longer lead times, making it a potential candidate for mid- to long-term predictions for both T2M and TP.

The T2M forecasts exhibited high accuracies in the first week across all regions (Table 1), with the highest accuracy

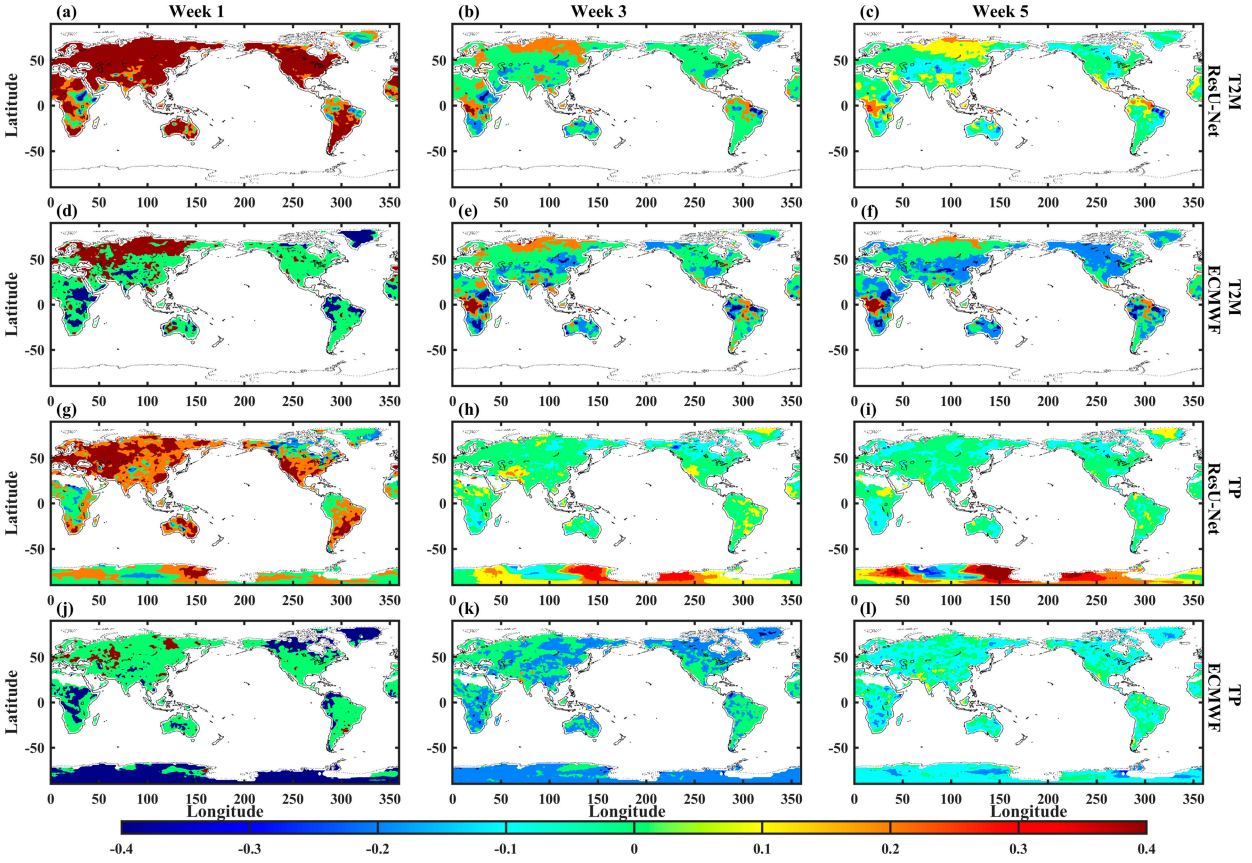

FIG. 3. Weekly RPSS for lead weeks 1, 3, and 5 in 2020. The T2M forecasts (a)–(c) from ResU-Net and (d)–(f) from ECMWF. The TP forecasts (g)–(i) from ResU-Net and (j)–(l) from ECMWF. Dry regions, such as the Sahara Desert, are excluded from the TP RPSS calculation.

observed in the NH. In week 1, the tropics had the lowest RPSS. There was a noticeable decrease in accuracy from weeks 1 to 2. Similarly, in the NH group, week 2 had the highest forecast accuracy, whereas the SH group had the lowest accuracy. Forecasting accuracies decreased significantly in week 3. Interestingly, the tropics showed a slight improvement over the NH, whereas the SH showed the least improvement. In week 4, the forecast accuracy showed a large

TABLE 1. RPSS of T2M and TP calculated based on ECMWF/U-Net/ResU-Net at different regions during the year 2020. Bold highlights indicate the best model.

	Time	NH	Tropics	SH	Global
T2M	Week 1	0.310/**0.531**/0.526	0.180/**0.395**/0.391	0.290/**0.504**/0.489	0.250/**0.473**/0.470
	Week 2	0.190/0.272/**0.278**	0.140/0.231/**0.232**	0.110/0.174/**0.176**	0.162/0.245/**0.252**
	Week 3	0.050/0.088/**0.094**	0.070/0.103/**0.110**	0.020/0.013/**0.039**	0.045/0.080/**0.088**
	Week 4	0.010/0.041/**0.043**	0.040/0.059/**0.065**	−0.030/−0.021/**−0.004**	0.014/0.037/**0.042**
	Weeks 3–4	0.053/**0.071**/0.067	0.068/0.092/**0.100**	−0.013/0.019/**0.032**	0.045/0.066/**0.067**
	Week 5	0.000/0.028/**0.033**	0.030/0.053/**0.057**	−0.030/−0.012/**0.002**	0.002/0.027/**0.034**
	Week 6	−0.010/0.018/**0.025**	0.030/0.046/**0.056**	−0.030/−0.031/**0.003**	−0.008/0.020/**0.031**
	Weeks 5–6	0.001/**0.016**/0.013	0.046/0.059/**0.068**	−0.047/−0.018/**0.023**	0.002/0.021/**0.026**
TP	Week 1	0.200/**0.310**/0.302	0.150/**0.288**/0.282	0.340/**0.355**/0.350	0.173/**0.290**/0.283
	Week 2	0.070/**0.101**/0.100	0.100/**0.148**/0.145	0.120/0.124/**0.126**	0.082/**0.117**/0.115
	Week 3	−0.020/0.016/**0.024**	0.020/0.057/**0.058**	0.010/**0.034**/0.031	−0.004/0.031/**0.035**
	Week 4	−0.020/0.008/**0.013**	0.010/0.034/**0.039**	−0.010/0.014/**0.016**	−0.009/0.016/**0.020**
	Weeks 3–4	−0.012/0.011/**0.016**	0.017/**0.055/0.055**	−0.003/**0.018**/0.005	−0.004/0.024/**0.026**
	Week 5	−0.020/0.007/**0.012**	0.000/0.027/**0.029**	0.000/0.012/**0.013**	−0.011/0.013/**0.016**
	Week 6	−0.020/0.004/**0.013**	−0.010/0.028/**0.031**	−0.010/**0.017**/0.009	−0.012/0.011/**0.017**
	Weeks 5–6	−0.023/0.003/**0.008**	−0.005/0.034/**0.036**	−0.014/0.006/**0.010**	−0.019/0.012/**0.015**

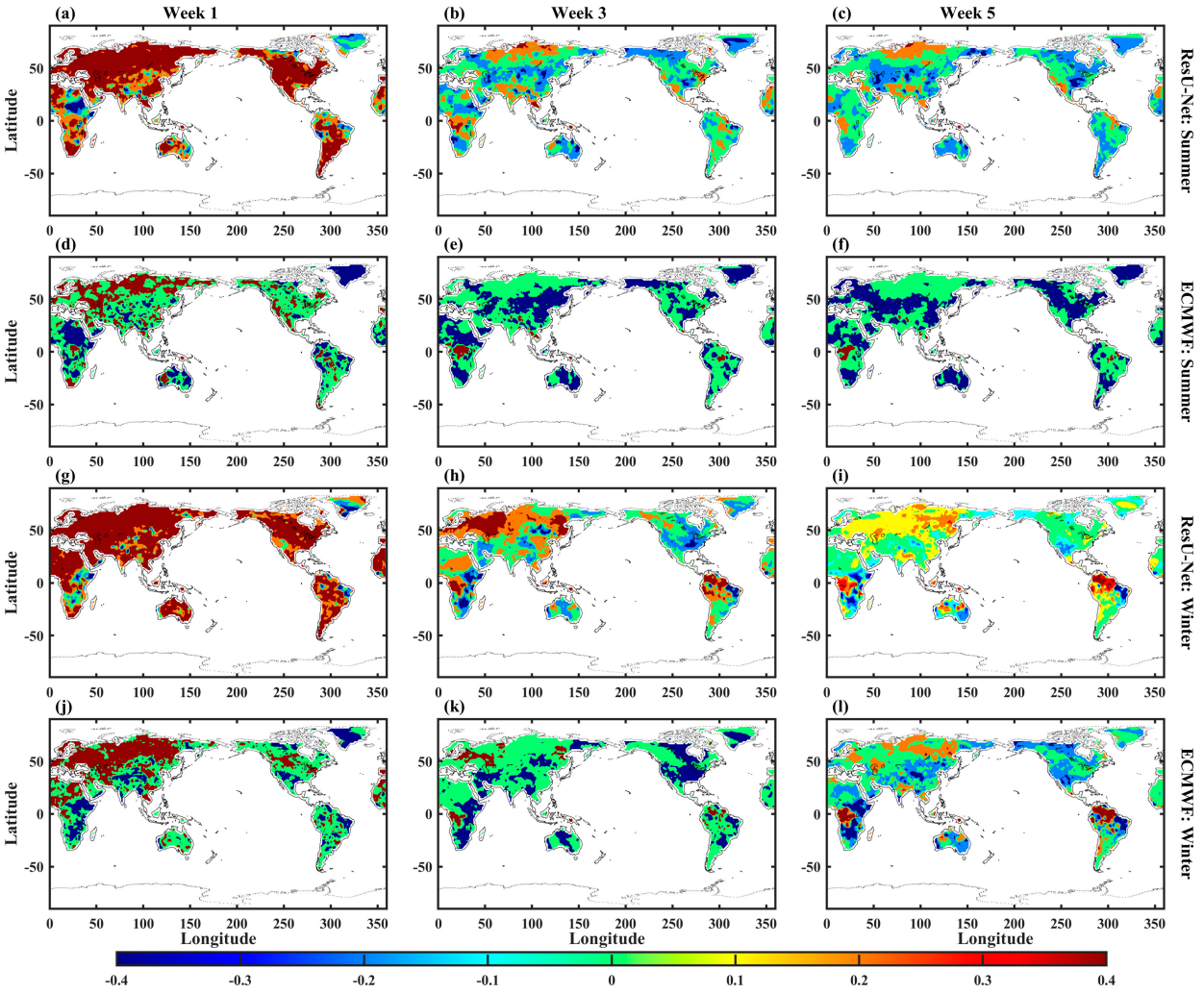

FIG. 4. Weekly RPSS for T2M during (a)–(f) boreal summer and (g)–(l) winter seasons in 2020. (a)–(c),(g)–(i) Forecasts from ResU-Net and (d)–(f),(j)–(l) forecasts from ECMWF.

decrease with negative values in the SH, indicating a poor performance compared to the climatology. The combination of weeks 3 and 4 showed a slight increase in accuracy compared with week 4 alone, especially in the tropics. In week 5, the tropics exhibited the highest RPSS values for T2M, indicating a relatively superior forecasting performance compared to the NH and SH. The NH showed moderate skill with slightly positive RPSS values. In contrast, the SH showed negative to near-zero values, suggesting that the model's performance was less reliable and potentially worse than the climatological forecasts in this region. In week 6, the forecasting skill for T2M decreased further across all regions. The tropics maintained the highest RPSS, although these values were lower than those at week 5. The NH showed a slight decrease in accuracy, with lower positive values than in the previous week. The SH continued to exhibit the lowest accuracy, with negative values indicating a poor forecasting performance. In weeks 5–6, the forecasting accuracy was low overall, with the tropics having the highest RPSS among all regions.

The global T2M forecast showed the highest skill in week 1, with RPSS values of 0.25 (ECMWF), 0.47 (U-Net), and 0.47 (ResU-Net). Forecast skill decreased with lead time, giving week 2 values of 0.16, 0.25, and 0.25 and week 3 values of 0.04, 0.08, and 0.09 for ECMWF, U-Net, and ResU-Net, respectively. During week 4, skill dropped further to 0.01, 0.04, and 0.04, and only marginal positive skill persisted through weeks 5 (0.00, 0.03, 0.03) and 6 (−0.01, 0.02, 0.03). The biweekly aggregates showed similar patterns, with RPSS values of 0.07 (ResU-Net) for weeks 3–4 and 0.03 (ResU-Net) for weeks 5–6, while ECMWF forecasts were near zero. The relative improvement of ResU-Net over ECMWF was approximately 60%–80% for weeks 1–3 and 30%–50% for weeks 4–6. This demonstrates that ResU-Net effectively decreases biases and maintains moderate skill even at extended subseasonal leads.

TP forecasts showed high accuracy in the first week, especially in the SH, whereas the tropics exhibited a slightly lower accuracy. In week 2, there was a significant decrease in

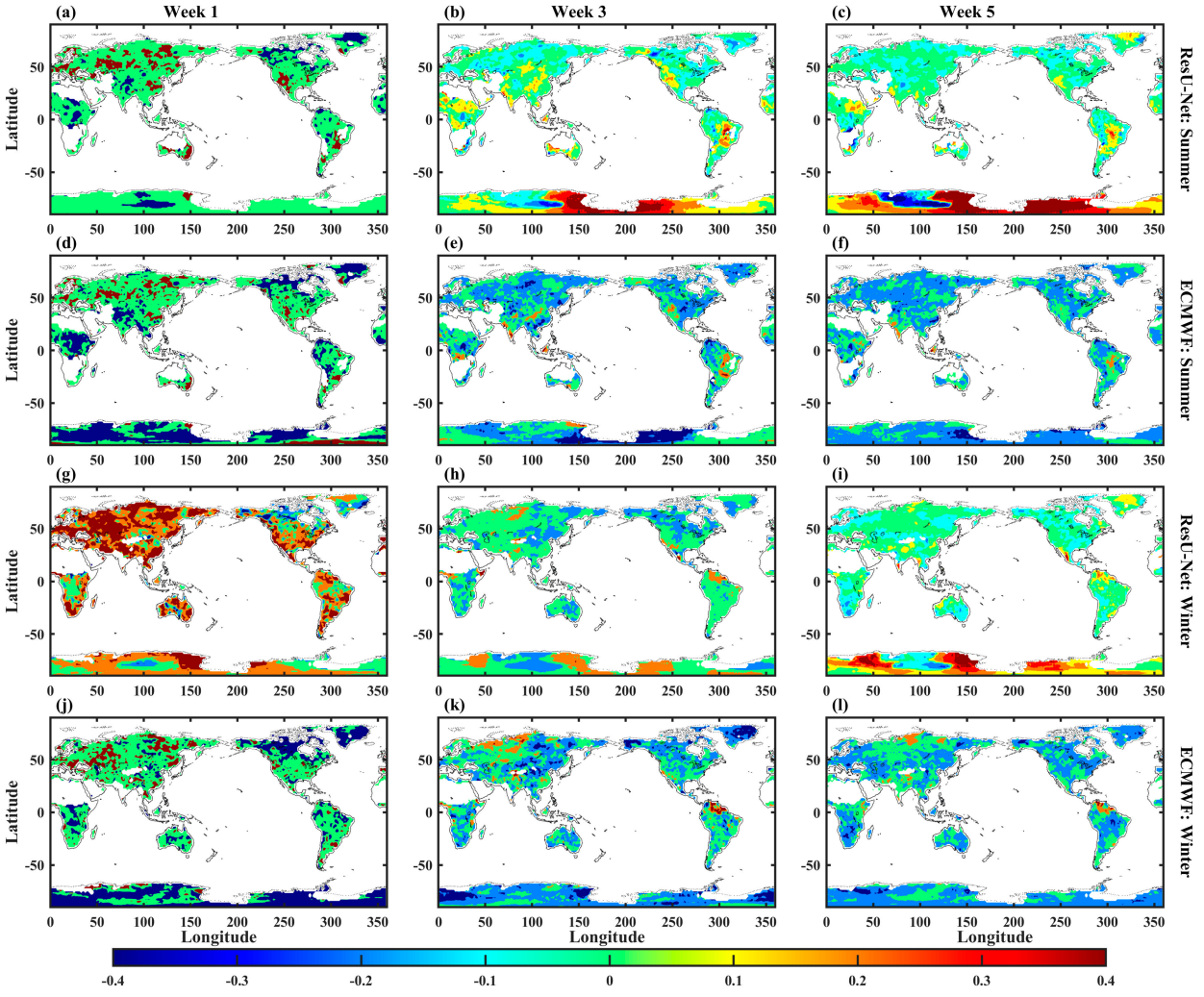

FIG. 5. Weekly RPSS for TP during (a)–(f) boreal summer and (g)–(l) winter seasons in 2020. (a)–(c),(g)–(i) Forecasts from ResU-Net and (d)–(f),(j)–(l) forecasts from ECMWF. Dry regions are excluded from the TP RPSS calculation.

forecast accuracy values, with the tropics maintaining higher accuracies than those of the NH and SH. In week 3, the forecasting accuracies were significantly low across all regions, with the tropics again showing a relatively superior performance. We observed extremely low accuracies in week 4, with slight improvements in the tropics. Weeks 5 and 6 showed a similar trend, with the tropics showing the highest skill among the three regions. NH demonstrated a considerably low accuracy and was barely positive, suggesting a limited predictive capability. The SH had small positive values, indicating some accuracy, but it was nonetheless relatively weak compared to that of the tropics.

For TP forecasts, the highest skill was scored at short leads, with week-1 RPSS values of 0.17 (ECMWF), 0.29 (U-Net), and 0.28 (ResU-Net). Skill declined by week 2 (0.08, 0.12, and 0.12) and week 3 (−0.00, 0.03, and 0.04), reaching nearly zero by week 4 (−0.01, 0.02, and 0.02). At longer leads, weeks 5 (−0.01, 0.01, and 0.02) and 6 (−0.01, 0.01, and 0.02) exhibited minimal skill. The biweekly averages confirmed this trend, with RPSS values around 0.03 (ResU-Net) for weeks 3–4 and 0.02 (ResU-Net) for weeks 5–6, compared with negative or near-zero ECMWF values. The corresponding fractional RPSS increase indicates a 50%–70% improvement for weeks 1–3 and a 20%–40% enhancement for weeks 4–6. These results suggest that ResU-Net provides better probabilistic forecasts than ECMWF.

Figure 4 illustrates the seasonal T2M forecast skill for weeks 1, 3, and 5, while the complete 6-week results are given in the appendix (Fig. A4). At shorter leads (week 1), ResU-Net outperforms ECMWF, showing higher skill across much of the Northern Hemisphere, the tropics, and parts of the Southern Hemisphere during both boreal summer (Fig. 4a) and winter (Fig. 4g). For lead weeks 3 and 5, skill decreases for both models, yet ResU-Net maintains higher RPSS values over southern Africa, South America, and northern Eurasia, particularly during boreal winter. The ECMWF model shows a rapid forecast skill degradation with near-zero or negative skill beyond week 2. During boreal winter, large-scale and

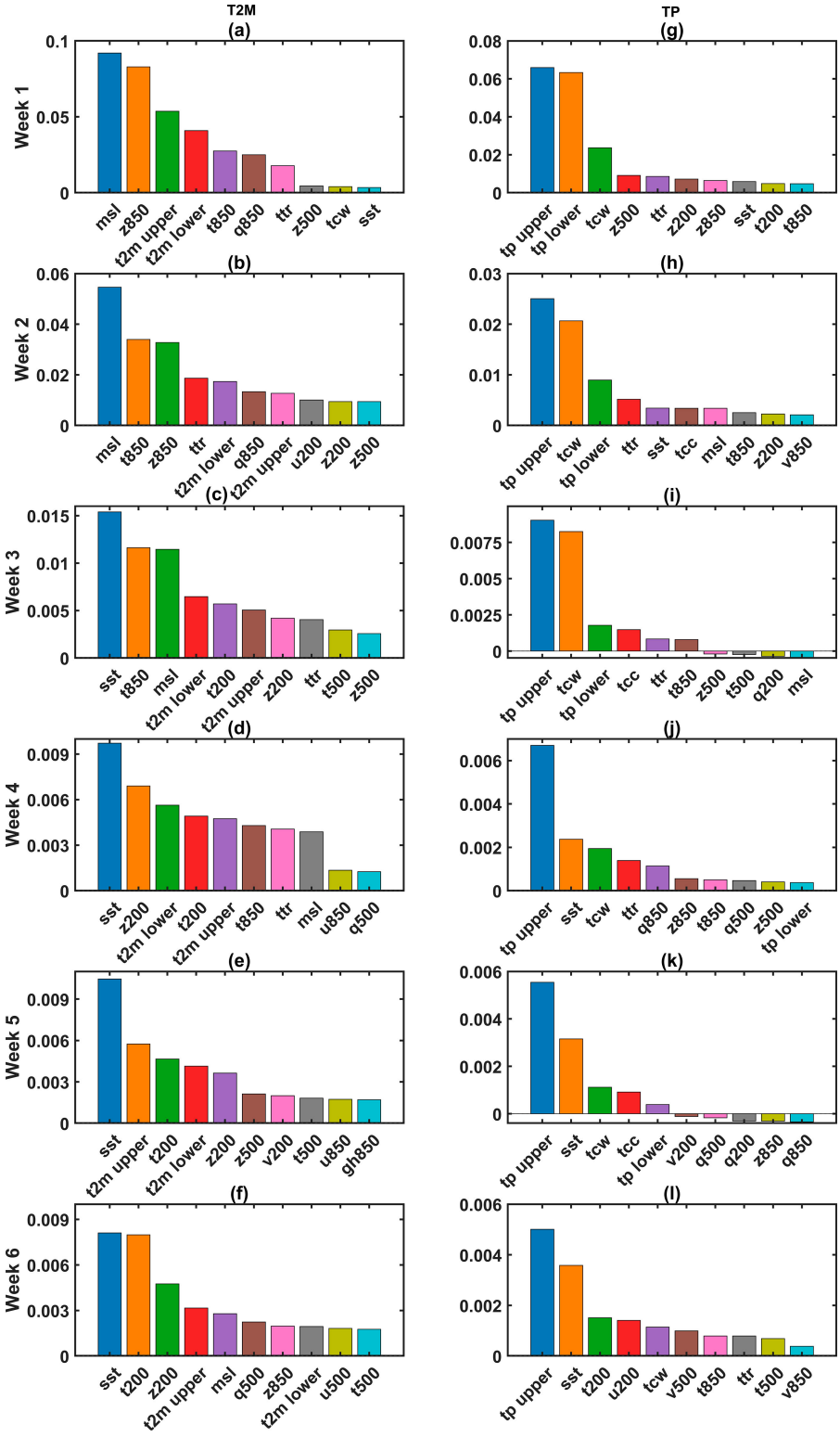

FIG. 6. Feature importance of the 10 most predictors in the ResU-Net model for the year 2020.

low-frequency climate modes such as the Arctic Oscillation (AO), North Atlantic Oscillation (NAO), and El Niño–Southern Oscillation (ENSO) modulate temperature anomalies over large spatial coherence and temporal persistence, creating favorable conditions for accurate T2M forecasts at S2S lead times (Johnson et al. 2019; Park et al. 2021).

Figure 5 depicts the seasonal forecast skill for TP for weeks 1, 3, and 5; the full results for all weeks are shown in the appendix (Fig. A5). In week 1, both models show positive RPSS in many regions, but ResU-Net exhibits stronger and more extensive skill. During boreal summer, ResU-Net captures precipitation variability associated with tropical convergence zones and the South Asian monsoon (Figs. A5b,c), effectively learn spatial–temporal patterns associated with monsoonal systems from the historical reforecast data. During boreal winter, it shows improved performance over North America, Europe, and the South Pacific convergence zone (Figs. A5n,o). As the lead time increases to week 5 (Figs. 5c,f,i,l), both models show reduced skill, but ResU-Net maintains coherent areas of positive RPSS, while ECMWF forecasts become near zero or negative over most regions.

b. *Predictor importance*

To investigate the relative importance of each predictor, permutation importance is calculated. Figure 6 shows the importance metrics for the top 10 predictors in the ResU-Net model for weeks 1–6. For T2M forecasts (Figs. 6a–f), the most important predictors during weeks 1 and 2 are MSL pressure, 850-hPa geopotential height (Z850), 850-hPa temperature (T850), and T2M upper and T2M lower, which represent the distance from the ensemble-mean forecast to the tercile category edges. The high importance of MSL, Z850, and T850 at short lead weeks highlights the critical role of synoptic-scale pressure patterns and lower-tropospheric thermal structure in controlling near-surface temperature variability (Vitart et al. 2017). For week 3, SST, T850, and MSL are the top important predictors, respectively. The most important predictors for T2M forecasts at lead times of weeks 4–6 are SST, 200-hPa geopotential height (Z200), and 200-hPa temperature (T200). This indicates that ocean–atmosphere coupling and large-scale variability associated with teleconnection patterns—such as ENSO, variability linked to the Arctic Oscillation, and fluctuations in the strength of the stratospheric polar vortex—can influence midlatitude temperature anomalies and thus strongly affect T2M forecasts at subseasonal time scales (Johnson et al. 2019; Mariotti et al. 2020).

For TP forecasts (Figs. 6g–i), the distance from the ensemble-mean forecast to the tercile category edges of precipitation (TP upper and TP lower) and TCW are key predictors for short lead weeks 1–3. These predictors are directly related to precipitation: TP upper and TP lower quantify the probabilistic proximity to wet or dry tercile thresholds, while TCW represents the integrated atmospheric moisture reservoir that governs condensation potential and rainfall formation. The importance of the ECMWF precipitation forecast as a predictor may reflect the persistence forecast skill of the model to capture large-scale moisture convergence and convective organization, which remain coherent over several weeks owing to the slow evolution of boundary layer humidity and oceanic forcing (Vitart et al. 2017; Johnson et al. 2019; Mariotti et al. 2020). At extended lead times (weeks 4–6), TP upper and SST become the most important predictors, indicating the growing influence of large-scale ocean–atmosphere coupling that organizes precipitation through phenomena such as ENSO (Mariotti et al. 2020).

4. Conclusions

This study developed a deep learning–based postprocessing framework (ResU-Net) for ECMWF subseasonal forecasts of 2-m temperature (T2M) and total precipitation (TP). Compared with the raw ECMWF forecasts and a baseline U-Net model, ResU-Net achieved higher ranked probability skill scores (RPSSs) globally and regionally, particularly during the first three forecast weeks. The largest relative improvement (60%–80%) occurred in early lead weeks, with ResU-Net maintaining positive but weaker skill through week 6.

These improvements demonstrate that ResU-Net algorithm effectively learns nonlinear relationships between predictands (T2M and TP) from large-scale predictors (e.g., SST, Z200, T850) and surface variables. The findings are consistent with recent advances in S2S forecast postprocessing (Vitart et al. 2022; Horat and Lerch 2024), where machine learning approaches enhanced ensemble calibration and bias reduction relative to purely physics-based systems.

The improved forecast skill demonstrates that data-driven postprocessing can serve as a robust probabilistic postprocessing tool for dynamical S2S forecasts. This approach could be integrated into operational ensemble systems to improve real-time forecasts for probabilistic early warning services.

In the future work, we will extend the postprocessing approach to the prediction of extreme precipitation and heat waves and to identify major sources of predictability (e.g., ENSO, MJO, stratospheric variability) using explainable AI methods to improve operational S2S forecasts. Additional experiments incorporating extended training datasets or physics-informed neural networks within the loss function may further enhance forecast skill and physical interpretability.

Acknowledgments. This work is supported by the Korea Meteorological Administration (KMA) Research and Development Program under Grant RS-2024-00403698 and Global–Learning & Academic research institution for Master's·PhD students, and Postdocs (G-LAMP) Program of the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. RS-2023-00301938).

Data availability statement. The data are available at the S2S AI challenge repository (https://s2s-ai-challenge.github.io/) and accessible through a dedicated CliMetLab plugin (https://github.com/ecmwf-lab/climetlab-s2s-ai-challenge). Python code for data processing and ResU-Net is available at https://github.com/Chalachew-github/resunet-s2s-forecast.

APPENDIX

Additional Figures

Figure A1 shows an example of the three TP categories: below normal, near normal, and above normal, based on observation, ECMWF, U-Net, and ResU-Net.

The global plot of RPSS values from U-Net for 2-m temperature (T2M) and total precipitation (TP) is given in Fig. A2.

The global plot of RPSS for T2M and TP is given in Fig. A3.

The global plot of RPSS for T2M during boreal summer and winter is given in Fig. A4.

The global plot of RPSS for TP during boreal summer and winter is given in Fig. A5.

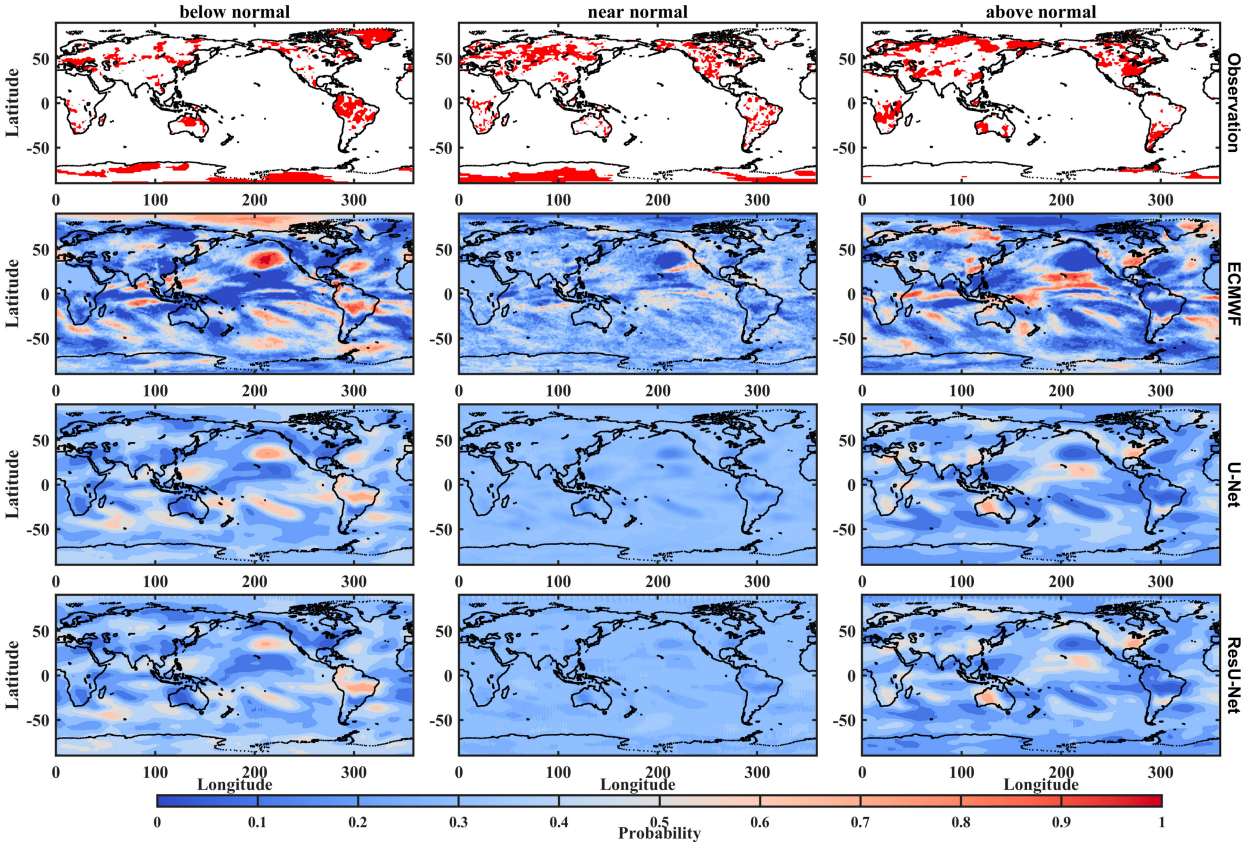

FIG. A1. Probability prediction of precipitation for lead week 2, issued on 2 Jan 2020.

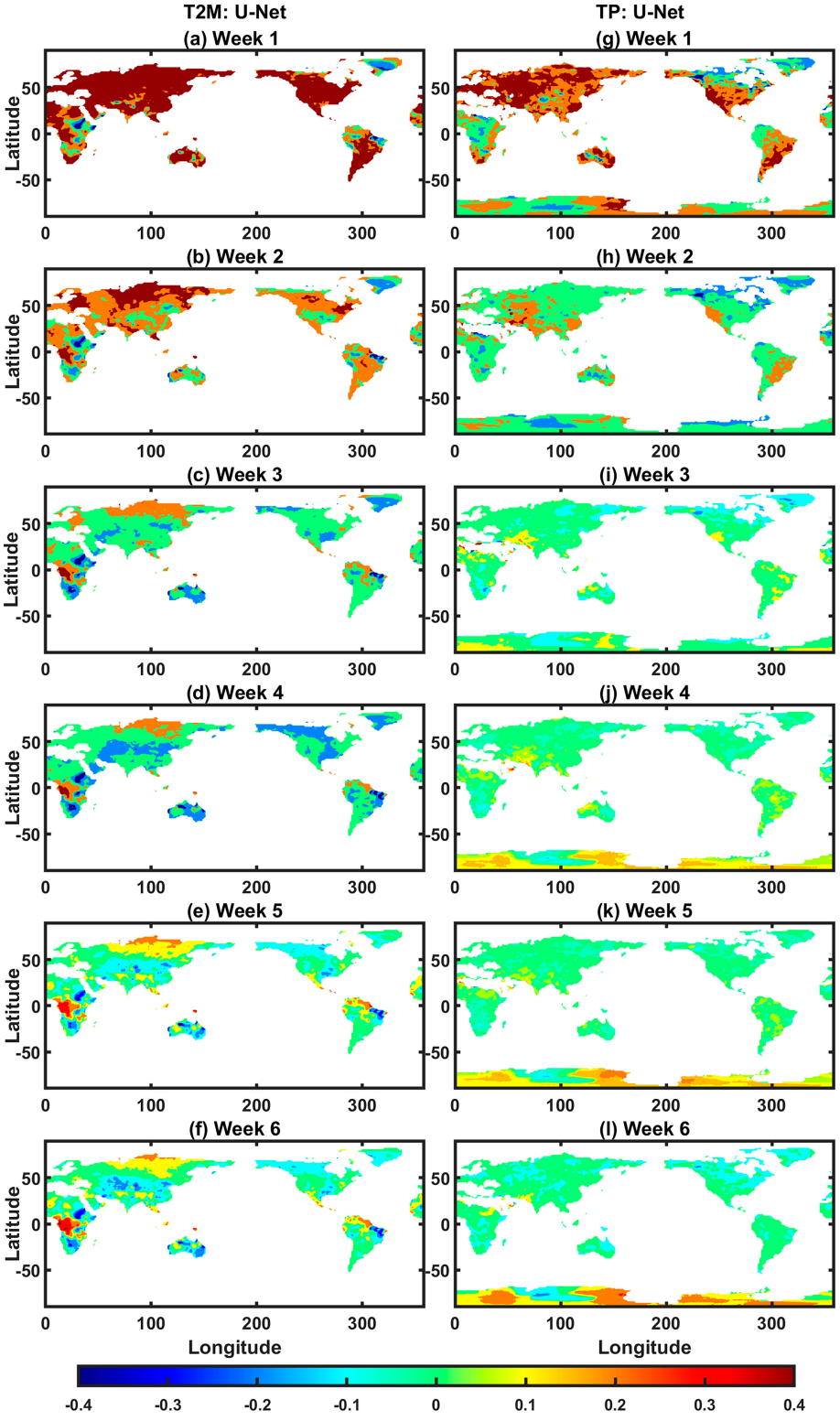

FIG. A2. U-Net-based weekly RPSS for (left) T2M and (right) TP of year 2020.

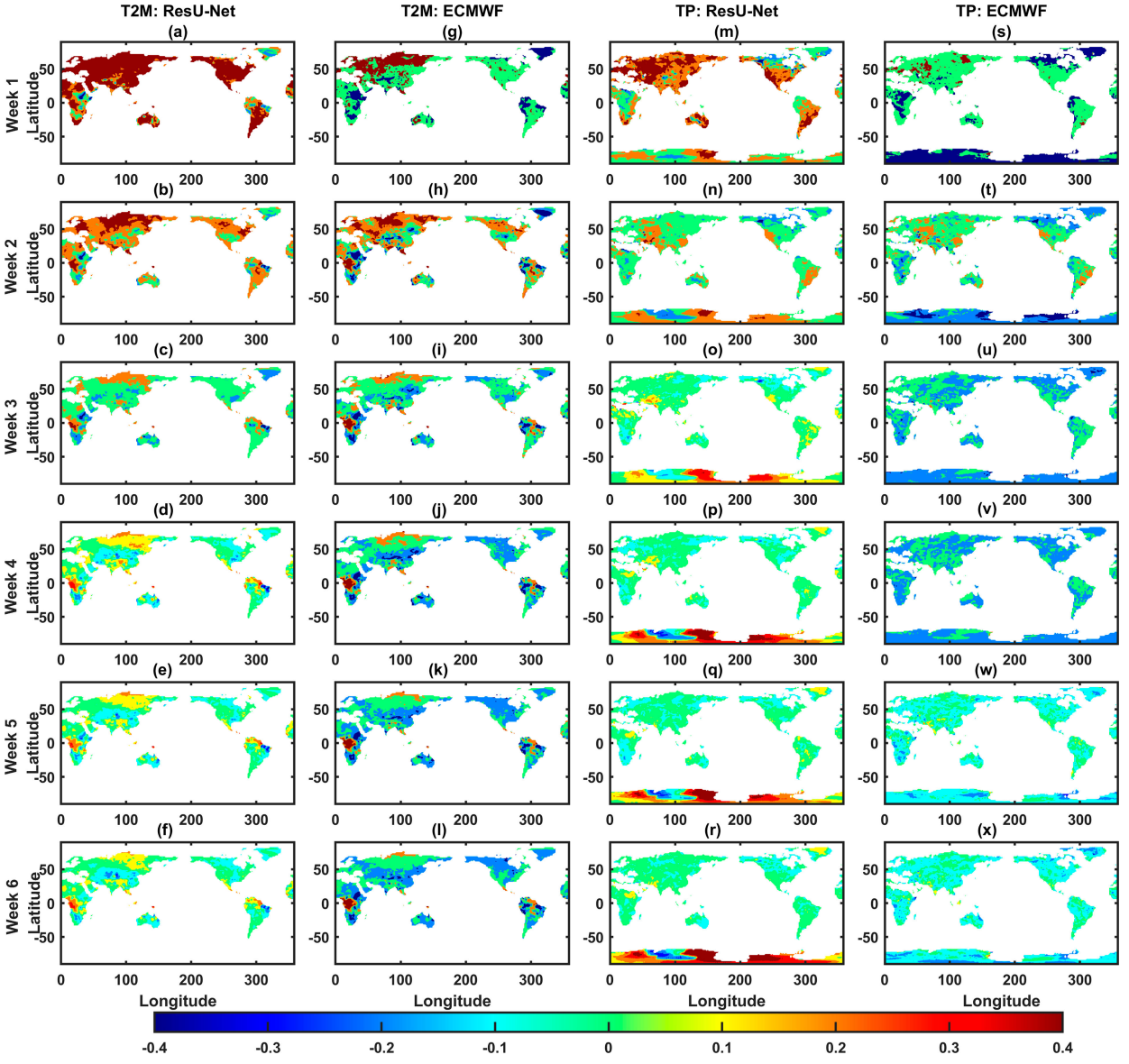

FIG. A3. Weekly RPSS for T2M and TP in 2020 from ResU-Net and ECMWF.

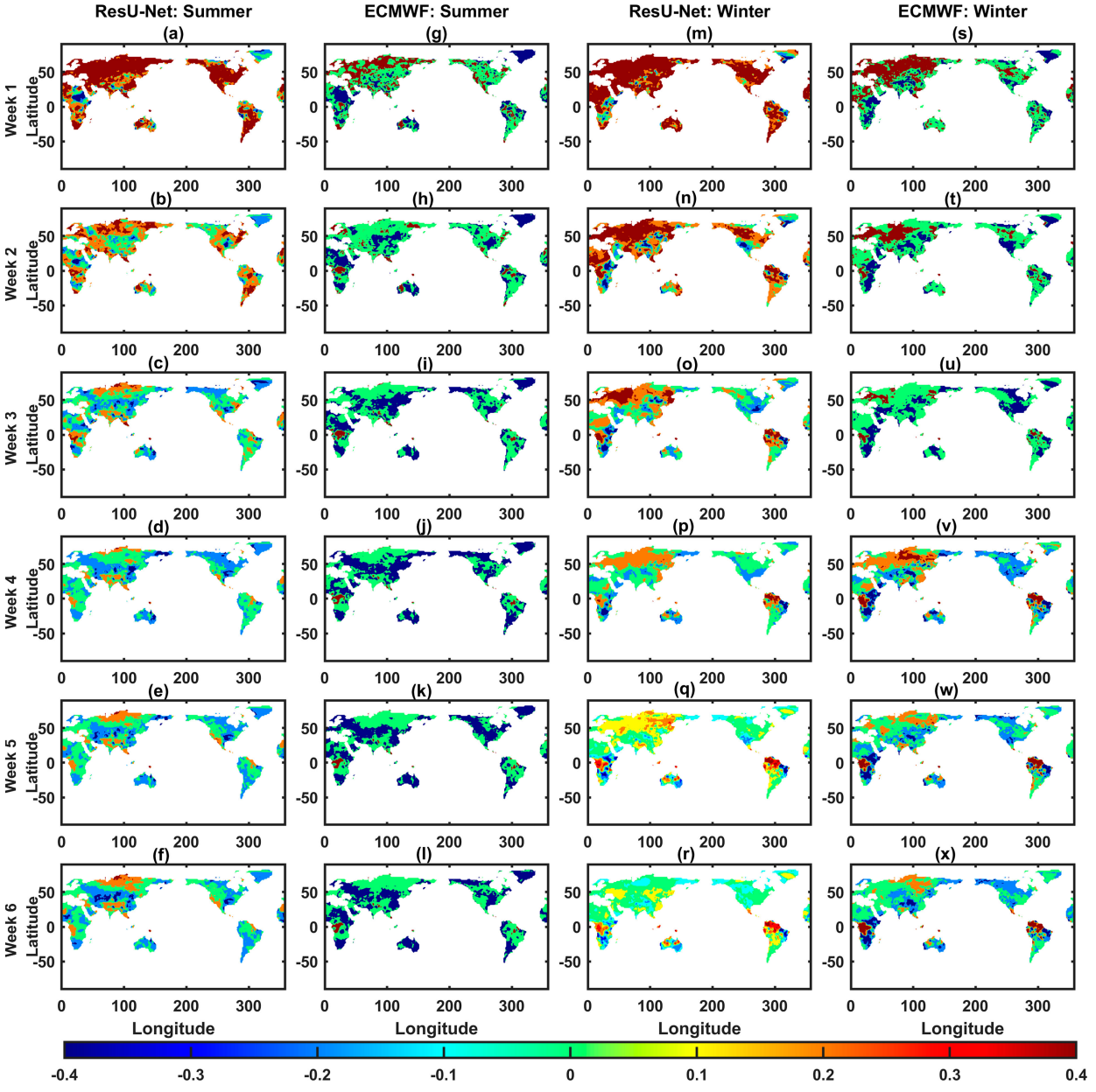

FIG. A4. Weekly RPSS for T2M during (a)–(l) boreal summer and (m)–(x) winter seasons in 2020.

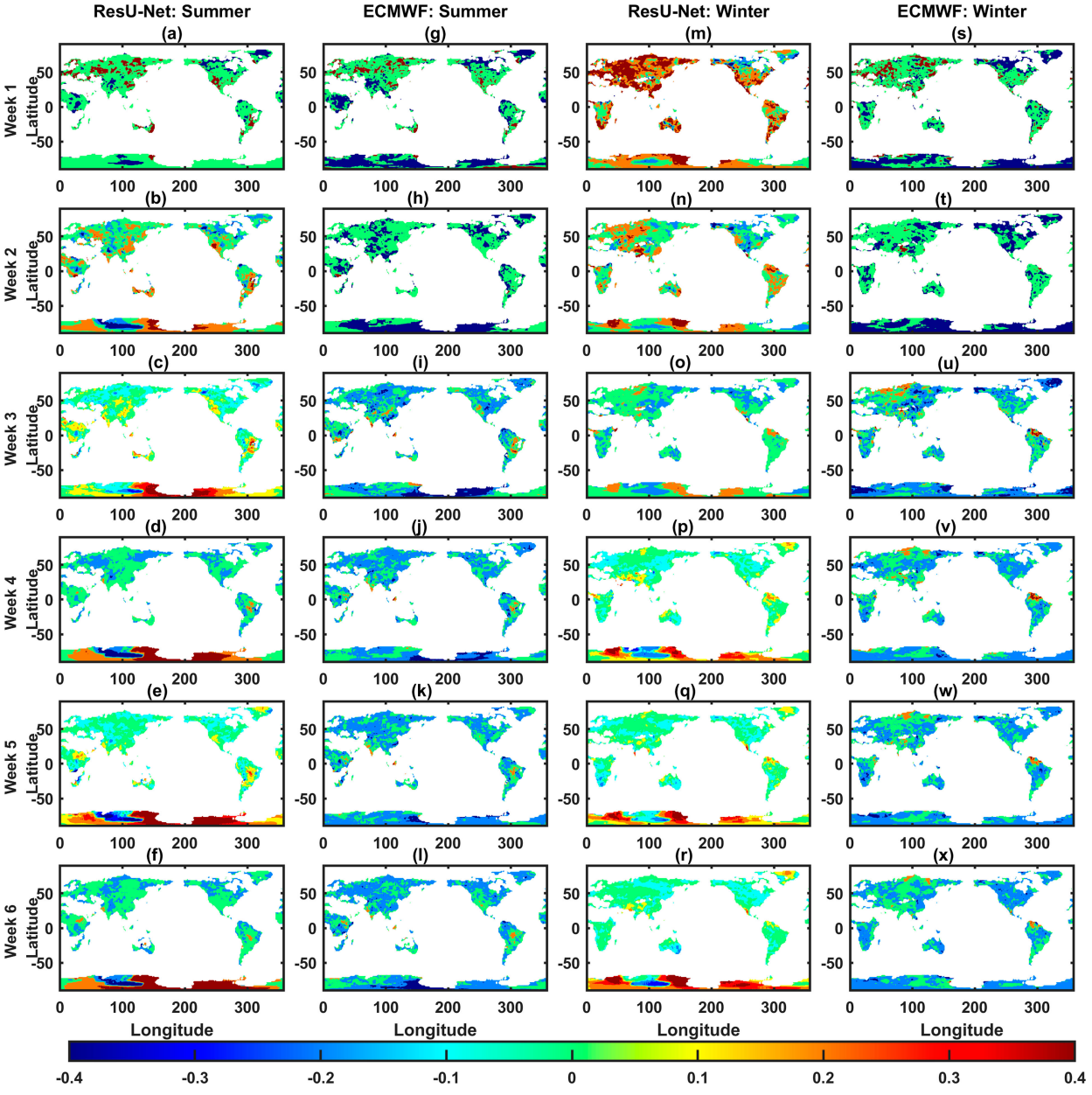

FIG. A5. Weekly RPSS for precipitation during (a)–(l) boreal summer and (m)–(x) winter seasons in 2020. Dry regions like the Sahara Desert are omitted from the RPSS.

REFERENCES

Bi, K., L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, 2023: Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, **619**, 533–538, https://doi.org/10.1038/s41586-023-06185-3.

Bonev, B., T. Kurth, C. Hundt, J. Pathak, M. Baust, K. Kashinath, and A. Anandkumar, 2023: Spherical Fourier Neural Operators: Learning stable dynamics on the sphere. arXiv, 2306.03838v1, https://doi.org/10.48550/arXiv.2306.03838.

Bridle, J. S., 1989: Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. *NIPS'89: Proc. Third Int. Conf. on Neural Information Processing Systems*, Cambridge, MA, MIT Press, 211–217, https://dl.acm.org/doi/10.5555/2969830.2969856.

Chapman, W. E., L. Delle Monache, S. Alessandrini, A. C. Subramanian, F. M. Ralph, S.-P. Xie, S. Lerch, and N. Hayatbini, 2022: Probabilistic predictions from deterministic atmospheric river forecasts with deep learning. *Mon. Wea. Rev.*, **150**, 215–234, https://doi.org/10.1175/MWR-D-21-0106.1.

Chen, L., and Coauthors, 2024: A machine learning model that outperforms conventional global subseasonal forecast models. *Nat. Commun.*, **15**, 6425, https://doi.org/10.1038/s41467-024-50714-1.

Clevert, D.-A., T. Unterthiner, and S. Hochreiter, 2016: Fast and accurate deep network learning by exponential linear units

(ELUs). arXiv, 1511.07289v5, https://doi.org/10.48550/arXiv.1511.07289.

Domeisen, D. I. V., and Coauthors, 2022: Advances in the subseasonal prediction of extreme events: Relevant case studies across the globe. *Bull. Amer. Meteor. Soc.*, **103**, E1473–E1501, https://doi.org/10.1175/BAMS-D-20-0221.1.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.*, **8**, 985–987, https://doi.org/10.1175/1520-0450(1969)008<0985:ASSFPF>2.0.CO;2.

Gröenquist, P., C. Yao, T. Ben-Nun, N. Dryden, P. Dueben, S. Li, and T. Hoefler, 2021: Deep learning for post-processing ensemble weather forecasts. *Philos. Trans. Roy. Soc.*, **A379**, 20200092, https://doi.org/10.1098/rsta.2020.0092.

Ham, Y.-G., J.-H. Kim, and J.-J. Luo, 2019: Deep learning for multi-year ENSO forecasts. *Nature*, **573**, 568–572, https://doi.org/10.1038/s41586-019-1559-7.

He, S., X. Li, T. DelSole, P. Ravikumar, and A. Banerjee, 2021: Sub-seasonal climate forecasting via machine learning: Challenges, analysis, and advances. *Proc. AAAI Conf. Artif. Intell.*, **35**, 169–177, https://doi.org/10.1609/aaai.v35i1.16090.

Horat, N., and S. Lerch, 2024: Deep learning for postprocessing global probabilistic forecasts on subseasonal time scales. *Mon. Wea. Rev.*, **152**, 667–687, https://doi.org/10.1175/MWR-D-23-0150.1.

Hu, W., M. Ghazvinian, W. E. Chapman, A. Sengupta, F. M. Ralph, and L. Delle Monache, 2023: Deep learning forecast uncertainty for precipitation over the western United States. *Mon. Wea. Rev*, **151**, 1367–1385, https://doi.org/10.1175/MWR-D-22-0268.1.

Johnson, S. J., and Coauthors, 2019: SEAS5: The new ECMWF seasonal forecast system. *Geosci. Model Dev.*, **12**, 1087–1117, https://doi.org/10.5194/gmd-12-1087-2019.

Kingma, D. P., and J. Ba, 2017: Adam: A method for stochastic optimization. arXiv, 1412.6980v9, https://doi.org/10.48550/arXiv.1412.6980.

Kochkov, D., and Coauthors, 2024: Neural general circulation models for weather and climate. *Nature*, **632**, 1060–1066, https://doi.org/10.1038/s41586-024-07744-y.

Lam, R., and Coauthors, 2023: Learning skillful medium-range global weather forecasting. *Science*, **382**, 1416–1421, https://doi.org/10.1126/science.adi2336.

Lang, S., and Coauthors, 2024: AIFS—ECMWF's data-driven forecasting system. arXiv, 2406.01465v2, https://doi.org/10.48550/arXiv.2406.01465.

Lyu, Y., S. Zhu, X. Zhi, Y. Ji, Y. Fan, and F. Dong, 2023: Improving Subseasonal-to-Seasonal prediction of summer extreme precipitation over southern China based on a deep learning method. *Geophys. Res. Lett.*, **50**, e2023GL106245, https://doi.org/10.1029/2023GL106245.

Mariotti, A., and Coauthors, 2020: Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bull. Amer. Meteor. Soc.*, **101**, E608–E625, https://doi.org/10.1175/BAMS-D-18-0326.1.

Merryfield, W. J., and Coauthors, 2020: Current and emerging developments in subseasonal to decadal prediction. *Bull. Amer. Meteor. Soc.*, **101**, E869–E896, https://doi.org/10.1175/BAMS-D-19-0037.1.

Messner, J. W., G. J. Mayr, and A. Zeileis, 2017: Nonhomogeneous boosting for predictor selection in ensemble postprocessing. *Mon. Wea. Rev.*, **145**, 137–147, https://doi.org/10.1175/MWR-D-16-0088.1.

Mouatadid, S., P. Orenstein, G. Flaspohler, J. Cohen, M. Oprescu, E. Fraenkel, and L. Mackey, 2023: Adaptive bias correction for improved subseasonal forecasting. *Nat. Commun.*, **14**, 3482, https://doi.org/10.1038/s41467-023-38874-y.

Murphy, A. H., 1971: A note on the ranked probability score. *J. Appl. Meteor.*, **10**, 155–156, https://doi.org/10.1175/1520-0450(1971)010<0155:ANOTRP>2.0.CO;2.

Park, H.-L., K.-H. Seo, B.-M. Kim, J.-Y. Kim, and S.-Y. S. Wang, 2021: Dominant wintertime surface air temperature modes in the Northern Hemisphere extratropics. *Climate Dyn.*, **56**, 687–698, https://doi.org/10.1007/s00382-020-05478-x.

Pathak, J., and Coauthors, 2022: FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. arXiv, 2202.11214v1, https://doi.org/10.48550/arXiv.2202.11214.

Pegion, K., and Coauthors, 2019: The Subseasonal Experiment (SubX): A multimodel subseasonal prediction experiment. *Bull. Amer. Meteor. Soc.*, **100**, 2043–2060, https://doi.org/10.1175/BAMS-D-18-0270.1.

Rasp, S., and S. Lerch, 2018: Neural networks for postprocessing ensemble weather forecasts. *Mon. Wea. Rev.*, **146**, 3885–3900, https://doi.org/10.1175/MWR-D-18-0187.1.

Robertson, A. W., F. Vitart, and S. J. Camargo, 2020: Subseasonal to seasonal prediction of weather to climate with application to tropical cyclones. *J. Geophys. Res. Atmos.*, **125**, e2018JD029375, https://doi.org/10.1029/2018JD029375.

Scheuerer, M., M. B. Switanek, R. P. Worsnop, and T. M. Hamill, 2020: Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over California. *Mon. Wea. Rev.*, **148**, 3489–3506, https://doi.org/10.1175/MWR-D-20-0096.1.

Taillardat, M., O. Mestre, M. Zamo, and P. Naveau, 2016: Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Mon. Wea. Rev.*, **144**, 2375–2393, https://doi.org/10.1175/MWR-D-15-0260.1.

Vannitsem, S., and Coauthors, 2021: Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bull. Amer. Meteor. Soc.*, **102**, E681–E699, https://doi.org/10.1175/BAMS-D-19-0308.1.

Veldkamp, S., K. Whan, S. Dirksen, and M. Schmeits, 2021: Statistical postprocessing of wind speed forecasts using convolutional neural networks. *Mon. Wea. Rev.*, **149**, 1141–1152, https://doi.org/10.1175/MWR-D-20-0219.1.

Vitart, F., and Coauthors, 2017: The Subseasonal to Seasonal (S2S) Prediction project database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1.

——, and Coauthors, 2022: Outcomes of the WMO prize challenge to improve subseasonal to seasonal predictions using artificial intelligence. *Bull. Amer. Meteor. Soc.*, **103**, E2878–E2886, https://doi.org/10.1175/BAMS-D-22-0046.1.

White, C. J., and Coauthors, 2017: Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteor. Appl.*, **24**, 315–325, https://doi.org/10.1002/met.1654.

——, and Coauthors, 2022: Advances in the application and utility of subseasonal-to-seasonal predictions. *Bull. Amer. Meteor. Soc.*, **103**, E1448–E1472, https://doi.org/10.1175/BAMS-D-20-0224.1.

Wilks, D. S., 2020: *Statistical Methods in the Atmospheric Sciences.* 4th ed. Elsevier, 818 pp.