Subseasonal to Seasonal (S2S) Prediction Algorithms Using Hybrid Machine Learning Techniques

HEE-SEUNG KIM,[a] SHANGLIN ZHOU,[b] ADAM BIENKOWSKI,[a] AND KRISHNA R. PATTIPATI[a]

[a] *Department of Electrical and Computer Engineering, University of Connecticut, Storrs, Connecticut*
[b] *Department of Computer Science and Engineering, University of Connecticut, Storrs, Connecticut*

(Manuscript received 6 December 2023, in final form 17 April 2025, accepted 26 May 2025)

ABSTRACT: Subseasonal to seasonal (S2S) prediction based on weather forecasts has received significant attention in recent years due to its potential application in various sectors, such as agriculture, energy, and water management. The challenges in S2S prediction stem from the complex and nonlinear interactions between atmospheric and oceanic processes, which can exhibit significant variability in short, medium, and long time scales. Traditional forecasting methods, such as physics-based numerical weather prediction (NWP) models, have limitations in capturing this variability and have shown limited skill in S2S prediction. However, recent advances in machine learning (ML) algorithms offer new opportunities for improving S2S predictions by enabling the identification of complex patterns and relationships in large and diverse datasets. This paper combines predictions from physics-based numerical weather prediction models with historical data to improve the forecast accuracy of temperature and precipitation averaged over forecast durations of 3–4 and 5–6 weeks using hybrid ML techniques. Specifically, a random forest classification model built for each of 23 regions of the world, as determined by a subject matter expert showed an overall 12.3% improvement in ranked probability skill score (RPSS) for temperature predictions and 4.2% for precipitation predictions for a forecast duration of 3–4 weeks; the improvement for forecast duration of 5–6 weeks was modest: 5.7% for temperature and 0.1% for precipitation.

SIGNIFICANCE STATEMENT: Machine learning algorithms have emerged as powerful tools in the weather forecast research, given large and diverse datasets. This study aims to investigate which features are useful for the subseasonal to seasonal prediction with a time scale of 2 weeks to 3 months based on hybrid machine learning techniques. Since the effects of features may vary depending on geographical and environmental conditions, we divide the global forecast area into 23 regions and improve the forecast accuracy in each region. This work has important implications for subseasonal to seasonal prediction and can be improved by using a wider variety of features available in the model for each region.

KEYWORDS: Forecast verification/skill; Statistical forecasting; Model evaluation/performance; Numerical weather prediction/forecasting; Classification; Machine learning

1. Introduction

Subseasonal to seasonal (S2S) prediction has emerged as a critical area of research in atmospheric science, with significant implications for weather forecasting, climate adaptation, and societal decision-making. S2S prediction refers to forecasting weather (e.g., temperature and precipitation) on a time scale of 2 weeks to 3 months, bridging the gap between short-range (i.e., up to about 10 days) and long-range (i.e., typically 3–6 months) weather forecasts (Vitart et al. 2017; Robertson et al. 2015). Interest in S2S prediction stems from its potential application in various sectors, such as agriculture, energy, and water management (White et al. 2017). S2S prediction is often referred to as a *predictability desert* due to its notorious difficulty in providing skillful predictions on subseasonal or monthly time scales (Hudson et al. 2011; Vitart et al. 2012).

The challenges of S2S prediction stem from the complex and nonlinear interactions between atmospheric and oceanic processes, which can exhibit significant variability over space and time due to local weather patterns as well as seasonal variations. Traditional forecasting methods, such as physics-based numerical weather prediction (NWP) models, have limitations in capturing this variability and have shown limited skill in S2S prediction (Bauer et al. 2015). For example, physics-based NWP models require precise initial conditions to generate accurate forecasts. However, uncertainties in initial observations, especially over data-sparse regions, can lead to structural errors that propagate over time, reducing forecast skill at the S2S scale (Bender and Ginis 2000; Koster et al. 2010). However, recent advances in machine learning (ML) algorithms offer potentially new opportunities for improving S2S predictions by enabling the identification of complex patterns and nonlinear relationships between atmospheric and oceanic processes in large and diverse datasets (Cohen et al. 2019; Hao et al. 2018) and combining these patterns with the predictions from numerical weather prediction models. ML algorithms can reduce uncertainty and enhance forecast skill by leveraging the diversity of individual models, capturing different aspects of the underlying processes, and providing more robust predictions (Weyn et al. 2021).

In this paper, we combine predictions from numerical weather prediction models with historical data to improve the

Corresponding author: Hee-Seung Kim, hee-seung.kim@uconn.edu

Publisher's Note: This article was revised on 5 August 2025 to correct a number of terms for style and consistency.

DOI: 10.1175/AIES-D-23-0108.1 e230108

forecast accuracy of temperature and precipitation globally averaged over a forecast duration of 3–4 and 5–6 weeks using ML techniques. Instead of predicting the actual values, we approach the problem as a categorical forecast, where the temperature and precipitation are classified into one of three categories (e.g., below normal, normal, or above normal). This classification-based approach enhances interpretability and aligns with decision-making needs in practical applications. Specifically, a random forest classification model built for each of the 23 regions of the world, as determined by a subject matter expert, showed an overall 12.3% improvement in ranked probability skill score (RPSS) for temperature predictions and 4.2% for precipitation predictions for a forecast duration of 3–4 weeks; the improvement for forecast duration of 5–6 weeks was modest: 5.7% for temperature and 0.1% for precipitation.

a. Weather forecast and its methods

A number of methods have been proposed to improve the accuracy and reliability of S2S predictions, including dynamical models, which are based on physical processes; statistical models, which are based on probabilistic relationships; and ML algorithms, which are based on learned relationships in the data.

Dynamical methods rely on NWP models, which simulate the entire climate system using physical equations governing the atmospheric and oceanic systems. Dynamical models, such as the Global Forecast System (GFS) and the European Centre for Medium-Range Weather Forecasts (ECMWF), are widely used for S2S prediction due to their ability to capture large-scale atmospheric and oceanic patterns (White et al. 2015). However, these models often suffer from biases and uncertainties, especially for subseasonal predictions.

Statistical methods use historical weather data to identify patterns and make predictions, typically through models like linear regression, moving averages, and autoregressive models. These methods rely on assumptions, such as linearity or specific data distributions, to interpret relationships (Breiman 2001b; Shmueli 2010). Since they are based on explicit mathematical relationships, statistical models are computationally efficient and easy to implement for structured datasets. However, their reliance on distributional and linear assumptions limits their ability to handle complex, nonlinear interactions between variables, especially in dynamic systems like the atmosphere and oceans. As a result, traditional statistical methods may struggle when confronted with new or unusual weather patterns. Advanced statistical models, such as multinomial logistic regression, quantile regression, and Bayesian model averaging, have been developed to improve prediction skills and reduce forecast errors for nonlinear relationships (Mendoza et al. 2015).

While statistical models can handle certain nonlinearities, their "shallow" architectures with limited number of model parameters constrain their ability to process complex patterns in large datasets. The advantages of ML-based methods, on the other hand, lie not only in their ability to represent nonlinear functions but also in their capacity to train large, deep neural networks. ML-based methods utilize extensive weight parameter spaces and sophisticated optimization algorithms to learn subtle correlations and interactions across variables, even in highly nonlinear and dynamic systems. This ability to model intricate dependencies and adapt to diverse scenarios underpins the exceptional performance of ML methods in applications, such as atmospheric and oceanic predictions (Bishop 2006; Schmidhuber 2015; Goodfellow et al. 2016).

Among the numerous ML algorithms for S2S prediction, wavelet transforms, random forest (RF), and support vector regression (SVR) are commonly utilized in the realm of hydroclimatic applications (Belayneh et al. 2014; Dikshit et al. 2020; Zeng et al. 2011). However, note that ML models require large amounts of training data and computational resources, and their "black box" nature can make the interpretation of model predictions more challenging when compared to statistical models (Jordan and Mitchell 2015; Bzdok et al. 2018).

While ML models offer a high capacity to learn complex and nonlinear relationships, it is essential to balance model complexity with generalization. Excessively complex models risk overfitting to noise in the training data, whereas overly simplistic models may fail to capture key predictive structures. The goal is to extract robust predictive signals from high-dimensional, noisy inputs that enable reliable forecasting in unseen periods (Bishop 2006; Shalev-Shwartz and Ben-David 2014). This trade-off, known as the bias–variance dilemma, highlights that while simple models may produce biased predictions, complex models can yield predictions with high variability. An effective forecasting model, therefore, strikes a careful balance between bias and variance.

b. Machine learning methods

Prior work on weather prediction using ML has demonstrated promising results in improving prediction skills and reducing errors compared to traditional statistical and dynamic models. Wu et al. proposed a two-layer stacking and blending ensemble algorithm for the prediction of daily reference evapotranspiration, i.e., the process by which water is transferred from the land to the atmosphere by evaporation from the soil and other surfaces and by transpiration from plants. The initial layer incorporated a range of methods including RF, SVR, multilayer perceptron (MLP) neural network, and k-nearest neighbors (kNNs). Notably, the stacking and blending of these models resulted in considerably improved predictive accuracy compared to the individual base models. Given the significant enhancement in performance demonstrated by this approach, it is strongly recommended for the accurate forecasting of reference evapotranspiration (Wu et al. 2021).

de Oliveira e Lucas et al. proposed a set of three convolutional neural networks (CNNs) to forecast time series data for reference evapotranspiration. Here, ensemble models were constructed by amalgamating these three individual CNNs. The outcome of these CNN ensembles was marked by predictions characterized by exceptional accuracy and remarkably low variance (de Oliveira e Lucas et al. 2020). The extreme gradient boosting (XGBoost) (Chen and Guestrin 2016) represents an enhanced iteration of the GB technique,

incorporating parallel preprocessing at the node level to enhance the computational speed compared to GB. Notably, XGBoost introduces a diverse range of regularization techniques aimed at mitigating the overfitting tendencies. Specifically, Chrit et al. presented a wind and turbulence prediction system designed for advanced air mobility application using a long short-term memory-based recurrent neural network (LSTM-RNN) model (Chrit and Majdi 2024). The system aims to support uncrewed aircraft system (UAS) integration into the National Airspace System (NAS) using ground-based wind data, such as wind speed, wind direction, wind gust, and eddy dissipation rate. Validated with airport and radiosonde data, this method outperforms MLP and XGBoost but struggles with lake-breeze predictions due to limited training data.

In recent years, there have been significant advancements in weather prediction models leveraging novel architectures and techniques (Chen et al. 2023; Rasp et al. 2024). Bi et al. proposed a Pangu-Weather algorithm, a deep learning–based system for fast and accurate global weather forecasting. By training deep neural networks with 43 years of hourly global weather data from ECMWF reanalysis data, Pangu-Weather achieves a spatial resolution comparable to the ECMWF Integrated Forecasting Systems. Pangu-Weather excels in short- to medium-range forecasts and supports various downstream scenarios, including extreme weather forecasting and large-member ensemble forecasts in real time (Bi et al. 2022).

Kurth et al. reported a Fourier Forecasting Neural Network (FourCastNet) method, a data-driven deep learning Earth system emulator designed to address the limitations of physics-based NWP methods, particularly in accuracy and computational efficiency. The predictions of the FourCastNet method enable accurate forecasts up to a week in advance and facilitate large ensembles to improve predictions of rare weather extremes (Kurth et al. 2023). Pathak et al. proposed a global data-driven high-resolution weather forecasting model based on adaptive Fourier neural operators (AFNOs). The model efficiently predicts weather patterns using spatiotemporal data, leveraging AFNO's ability to capture global dependencies, outperforming traditional methods in accuracy and computation time (Pathak et al. 2022).

The graph neural network (GraphCast) (Lam et al. 2023) method utilized a significant advance in accurate and efficient global medium-range weather forecasting, leveraging machine learning to model complex dynamical systems effectively. This algorithm predicts hundreds of weather variables over 10 days at a 0.25° resolution globally in under 1 min. GraphCast outperforms operational deterministic systems on 90% of 1380 verification targets, supporting better prediction of severe weather events such as tropical cyclones, atmospheric rivers, and extreme temperatures.

Price et al. presented a machine learning–based generative model for global medium-range probabilistic weather forecasting called GenCast. This algorithm employs a diffusion model to generate ensembles by sampling from the joint distribution of future weather trajectories (Price et al. 2023). Verbitski et al. discussed Amazon Aurora, a cloud-native relational database designed for high throughput, focusing on performance, scalability, and fault tolerance through innovative storage and recovery mechanisms (Verbitski et al. 2017).

Kashinath et al. explore the use of physics-informed ML (PIML) to enhance weather and climate modeling. This paper presents 10 case studies where PIML is applied to areas like extreme weather prediction and climate dynamics. By integrating physical principles with machine learning, PIML reduces data requirements while enhancing computational efficiency and predictive capabilities, addressing limitations of traditional models (Kashinath et al. 2021).

In the context of S2S modeling challenge organized by the World Meteorological Organization (WMO) in 2021, the Computer Research Institute of Montreal S2S (CRIMS2S) team proposed an opportunistic mixture model. It consists of a weighted multimodel ensemble based on five predictions: ECMWF, Environment and Climate Change Canada (ECCC), and NCEP forecasts, each postprocessed using ensemble model output statistics (EMOS). CRIMS2S team applied a prediction based on a CNN to the ECMWF forecasts (Vitart et al. 2022; Gneiting et al. 2005). We seek to improve on these predictions in this paper.

c. *Contribution and organization of the paper*

As an extension of our work at the WMO S2S artificial intelligence (AI) challenge in 2021, we present the S2S prediction algorithms using hybrid machine learning methods based on the combination of expert-defined geographic regions and the use of explainable AI techniques, such as Shapley additive explanation (SHAP) values and partial dependence plot (PDP) for regional and seasonal interpretation.

The integration of expert-defined geographic regions, as opposed to data-driven clustering methods commonly used in machine learning, further enhances the novelty of our approach. By incorporating expert knowledge of climate dynamics into our machine learning models, we not only improve prediction accuracy but also increase the interpretability of the results, making the model more transparent and trustworthy for stakeholders.

Additionally, our use of SHAP values and PDP to interpret the contributions of different variables across expert-defined geographic regions distinguishes our work from existing models that typically apply these techniques to more generic or globally aggregated data. This approach enables us to offer a detailed interpretation of the impact of key climate factors across different regions and time periods, thereby improving the interpretability and transparency of our machine learning models.

The paper is organized as follows. In section 2, we provide a brief description of the S2S dataset, including external observable features and the physics-based ECMWF forecasts that form the inputs to our ML algorithms. We provide metrics to measure and analyze the performance of the proposed model in section 3. Section 4 shows the probabilistic forecast of temperature for forecast 3–4 weeks (days 15–28). We discuss the S2S prediction results for precipitation for forecast 3–4 weeks and 5–6 weeks (days 29–42) in section 5. Last, we conclude the paper and discuss potential avenues for future work in section 6.

TABLE 1. Derived features.

Features at location x	Description
N_d	Observations at x for the past N_d days
N_y	Observations at x on the same day of the year for the past N_y years
Bias statistics	Hindcast realizations (avg, std)
Biweekly statistics	Biweekly historical statistics (avg, std, skw, krt, med)
tp observation	Total precipitation observation
tp bias	Average of hindcast realizations of total precipitation

2. Datasets and their characteristics

In this section, we describe the datasets used and the ancillary parameters that we experimented with as features in our models, including the ECMWF forecasts and observation sequences, as well as the measured and ECMWF-forecasted external variables. For all datasets, the years 2000–18 were used for training, 2019 was used for validation, and 2020 was used for testing.

a. Derived features from the S2S observations

We process the S2S observation sequences to extract features to help improve the accuracy of temperature and precipitation forecasts. Since the observation sequences are chronological, the first two derived features are observations for the past N_d days and observations on the same day of the year as the forecast day for the past N_y years. The hyperparameters N_d and N_y vary depending on models used for different forecast lead times (14 and 28 days) and for different response variables or quantities of interest, e.g., temperature and precipitation. Biweekly statistics, such as biweekly mean (avg), standard deviation (std), median (med), skewness (skw), and kurtosis (krt), are computed from the observation sequences and are used as features to help improve the prediction accuracy as well. Table 1 shows the derived features used in the models.

b. Additional observable parameters as features

Based on recommendations from weather experts, additional observable parameters are included as features in our models to help improve the prediction accuracy. A summary of the external features investigated is included in Table 2.

El Niño (Dai and Wigley 2000; Ropelewski and Halpert 1986) is the area-averaged sea surface temperature. We use two El Niño indices for two distinct spatial ranges, El Niño-1.2 and El Niño-3.4, from the National Oceanic and Atmospheric Administration (NOAA) Physical Sciences Laboratory. El Niño-1.2 is from 0°–10°S to 90°–80°W, and El Niño-3.4 is from 5°S–5°N to 170°–120°W. Both of the El Niño indices are available monthly and have the same values for all the locations and are available online (https://psl.noaa.gov/gcos_wgsp/Timeseries/).

The Pacific decadal oscillation (PDO) (Mantua and Hare 2002) is a long-lived El Niño–like pattern of Pacific climate variability. Widespread variations in the Pacific basin and the North American climate can mark the extremes in the PDO pattern. We use PDO data from NOAA's National Centers for Environmental Information (NCEI), which is based on NOAA's Extended Reconstructed SST (ERSST) (https://www.ncei.noaa.gov/access/monitoring/pdo/).

The North Atlantic Oscillation (NAO) (Barnston and Livezey 1987) is based on the difference in the surface sea level pressure between the subtropical (Azores) high and the subpolar low. Basinwide changes in the intensity and location of the North Atlantic jet stream and storm track, large-scale modulations of the normal patterns of zonal and meridional heat, and moisture transport would impact the value of NAO. In this case, NAO could in turn impact the changes in temperature and precipitation patterns. The NAO dataset that we use is from NOAA's NCEI (https://www.ncei.noaa.gov/access/monitoring/nao/).

Quasi-biennial oscillation (QBO) (Baldwin et al. 2001) is collected through the daily wind observations at selected stations near the equator. In the data, the monthly mean zonal wind components were calculated at 7 levels, which are 70, 50, 40, 30, 20, 15, and 10 hPa. We select different levels of the QBO in different models as they provided different degrees of impact. The QBO dataset we use is from the Free University of Berlin (https://www.geo.fu-berlin.de/en/met/ag/strat/produkte/qbo/index.html).

Sunspot cycles (Meehl et al. 2009) have long been correlated with weather patterns. Although the science behind the correlations is still unresolved, this feature shows that these small fluctuations in solar output can produce unexpectedly large responses in tropical precipitation, sea surface temperature, and cloud cover. In general, sea surface temperatures, soil moisture, soil temperature, ice cover, and snow cover all reflect long-term trends in temperature and precipitation. As such, anomalies in these features likely provide useful information for seasonal forecasts. The dataset we use is from NOAA's Space Weather Prediction Center (SWPC) (https://www.swpc.noaa.gov/products/solar-cycle-progression).

TABLE 2. External variables.

Name	Description	Format
El Niño	Area-averaged sea surface temperature. Two variables	Monthly data for all locations
Solar	Sunspot cycles. Two variables: original value and smoothed value	Monthly data for all locations
NAO	North Atlantic Oscillation	Monthly data for all locations
PDO	Pacific decadal oscillation	Monthly data for all locations
GL	Great Lakes ice cover	Daily data for all locations
QBO	Quasi-biennial oscillation	Monthly data for all locations

TABLE 3. ECMWF forecast datasets.

Name	Description	Format	Unit
rsn	Snow density	Daily average for all locations	kg m^{-3}
sm100	Soil moisture top 100 cm	Daily average for all locations	kg m^{-3}
sm20	Soil moisture top 20 cm	Daily average for all locations	kg m^{-3}
st20	Soil temperature top 100 cm	Daily average for all locations	K
msl	Mean sea level pressure	Instantaneous once per day	Pa
tcc	Total cloud cover	Daily average for all locations	%
tcw	Total column water	Daily average for all locations	kg m^{-2}
ci	Sea ice cover	Daily average for all locations	Proportion
gh	Geopotential height	Instantaneous once per day at 10 pressure levels	gpm
Wind	U velocity u and V velocity v	Instantaneous once per day at 10 pressure levels	m s^{-1}
q	Specific humidity	Instantaneous once per day at 7 pressure levels	kg kg^{-1}

c. *ECMWF forecast datasets*

The ECMWF also provides many additional forecast parameters that can help in the prediction of temperature and precipitation. The ECMWF hindcast model[1] begins by data assimilation with realistic estimates of the weather parameters derived from observations, subsequently making iterative weather predictions for a predefined extended duration devoid of boundary constraints. This process restarts on each Monday and Thursday for the time span of 1995–2016, effectively projecting the forthcoming 46-day weather evolution using an ensemble of 11 members. Notably, the model is coupled with an ocean model while excluding the sea ice model from its configuration. Over the validation period, a total of 1869 hindcast experiments were conducted (Vitart et al. 2017).

Two kinds of features are included in Table 3. The first group is comprised of single-level parameters: mean sea level pressure (msl), sea ice cover (ci), snow density (rsn), soil moisture in the top 100 cm (sm100), soil moisture in the top 20 cm (sm20), soil temperature in the top 20 cm (st20), total cloud cover (tcc), and total column water (tcw). Among these parameters, msl is collected once a day, and all the others are collected as a daily average value. For parameters that are collected as daily average values, we preprocess them in a manner similar to the preprocessing of the temperature as a biweekly average. While for parameters that are collected once a day, they are preprocessed in a manner similar to precipitation, which is aggregated as a biweekly difference. The second group is pressure-level parameters: geopotential height (gh), U velocity u, V velocity v, and specific humidity q. All these parameters are collected once a day, gh, u, and v are available at 10 levels of pressure, which are at 1000, 925, 850, 700, 500, 300, 200, 100, 50, and 10 hPa, while q is available at 7 levels of pressure, which are at 1000, 925, 850, 700, 500, 300, and 200 hPa.

[1] A hindcast refers to the process of using a numerical weather prediction model to simulate past weather or climate conditions. By starting with observed data as initial conditions, the model reproduces historical scenarios over a predefined period. It is commonly used in climate science to evaluate the model's accuracy by comparing its outputs with the actual observed data.

d. *ECMWF forecast uncertainty*

We discuss the uncertainties and limitations of the ECMWF forecasts (Epstein 1969b; Doblas-Reyes et al. 2013; Palmer and Anderson 1994). ECMWF forecasts depend on data from satellites, ground-based sensors, and weather balloons, but these data have limitations, including coverage gaps in areas like oceans and less-populated land regions. Additionally, measurement errors can spread through the models, leading to greater uncertainty, especially in regions with sparse data. This results in imperfect initial conditions and contributes to forecast inaccuracies (Lorenc 1986; Navon 2009).

The chaotic nature of the atmosphere, known as the "butterfly effect," means small differences in initial conditions can cause significantly different weather patterns over time. Even with high-resolution data used by the ECMWF, it is impossible to eliminate all errors, leading to increasing forecast uncertainty, especially for long-term predictions (Thompson 1957; Buizza 2002; Slingo and Palmer 2011). In addition, ECMWF models use parameterizations to approximate complex atmospheric processes like cloud formation and turbulence, which cannot be fully obtained at smaller scales. These simplifications, while necessary for computational efficiency, introduce limitations that affect forecast accuracy (Arakawa 2004; Ehrendorfer 1997; Dawson and Palmer 2015).

To address these inherent uncertainties, ECMWF employs ensemble forecasting, generating multiple simulations with slightly different initial conditions. This probabilistic approach provides a range of possible outcomes and helps quantify forecast uncertainty, although uncertainty inevitably grows over time. Efforts to improve forecasting techniques, such as integrating advanced data assimilation and machine learning, may help reduce these uncertainties in the future.

3. Evaluation metrics

a. *Ranked probability skill score*

Root-mean-square error (RMSE) (Chai and Draxler 2014; Willmott and Matsuura 2005) and mean bias (Li et al. 2019; Han et al. 2023) are more suited for deterministic forecasts, which provide a single predicted value to compare with observed values. RMSE measures the average error, while mean bias indicates if predictions consistently overestimate or underestimate

the actual values. However, both metrics ignore the probabilistic nature of forecasts and fail to capture the range of possible outcomes, which is crucial for S2S predictions.

RPSS is a relative accuracy metric commonly used in weather and climate forecasting to evaluate the performance of probabilistic predictions. The RPSS measures the improvement in the forecast skill of a model relative to a reference model that is typically a climatological forecast or a persistence forecast (Epstein 1969a; Murphy 1971; Weigel et al. 2007). The RPSS takes into account the entire forecast distribution rather than just the deterministic forecast value, making it a useful measure for assessing the performance of probabilistic forecasts, such as ensemble forecasts.

The ranked probability score (RPS) is a squared metric that assesses forecast performance by contrasting the cumulative density function (CDF) of a probabilistic forecast with the CDF of the corresponding observation (Epstein 1969a; Murphy 1971). The RPSS can be obtained by measuring RPS for the model and RPS for the climatological forecast, denoted as $\mathrm{RPS_{cl}}$ and combining them via

$$\mathrm{RPSS} = 1 - \frac{\langle \mathrm{RPS} \rangle}{\langle \mathrm{RPS_{cl}} \rangle} = 1 - \frac{\left\langle \sum_{m=1}^{M} \left[\left(\sum_{k=1}^{m} y_k \right) - \left(\sum_{k=1}^{m} o_k \right) \right]^2 \right\rangle}{\left\langle \sum_{m=1}^{M} \left[\left(\sum_{k=1}^{m} p_k \right) - \left(\sum_{k=1}^{m} o_k \right) \right]^2 \right\rangle}, \tag{1}$$

where M is the number of forecast categories (3 for terciles), y_k is the predicted probability for category k, o_k is the observed category (1 if the observation is in category k else 0), and p_k is the probability of category k based on climatography. The $\langle \cdot \rangle$ denotes the average of scores. From this equation, we can see that assuming we have equal amounts of samples in each category, then the best we can do with a constant y_k is $y_1 = y_2 = y_3 = 1/3$, given when M is 3.

b. Shapley additive explanation value

In machine learning, SHAP value (Lundberg and Lee 2017; Staniak and Biecek 2019) is a concept borrowed from cooperative game theory that assigns a value to each feature indicating its contribution to the prediction of a model. SHAP values provide transparent and interpretable insights into models, especially when dealing with complex, nonlinear interactions between variables. By identifying the influence of features, SHAP values play an important role in enhancing forecast accuracy and in evaluating uncertainty within S2S predictions. The SHAP values are computed on training data to determine the importance of features across a range of predictions. Each data point in a SHAP summary plot corresponds to a SHAP value for an individual feature for a given instance, showing how the prediction is influenced by that feature. Specifically, a positive SHAP value indicates that the feature contributes to increasing the predicted output relative to a baseline, which is defined as the expected value of the model output over the training data. Conversely, a negative SHAP value indicates that the feature contributes to decreasing the prediction relative to this baseline (Lundberg and Lee 2017; Lundberg et al. 2020; Christoph 2020).

SHAP values offer distinct advantages over traditional methods, such as backward selection, in feature importance analysis. While backward selection requires retraining multiple models to assess a feature's impact, SHAP computes the feature's importance from a single model, reducing the computational cost. Additionally, SHAP values provide both global and local interpretability, explaining feature contributions to the overall model predictions and for individual instances, unlike backward selection, which focuses only on the global importance of a feature (Wang et al. 2024; Marcílio and Eler 2020; Lee et al. 2023).

The SHAP value ϕ_{x_s} for a particular feature x_s in a model $f(x)$ determines the average marginal contribution by comparing the model's predictions with and without x_s from a subset S of the input feature set $x = \{x_1, x_2, \ldots, x_N\}$ excluding x_s given by

$$\phi_{x_s} = \sum_{S \subseteq x \{x_s\}} \frac{|S|!(N - |S| - 1)!}{N!} [f(S \cup \{x_s\}) - f(S)], \tag{2}$$

where $S \subseteq x \backslash \{x_s\}$ represents all possible subsets of features excluding x_s and N is the total number of features in the model. The terms $f(S \cup \{x_s\})$ and $f(S)$ are the model predictions using the features in S including and excluding x_s, respectively. Given $|S|$ is the size of the subset S and the ! symbol is the factorial operation, the term $|S|!(N - |S| - 1)!/N!$ ensures a fair weighting for the contribution of each subset since the contribution of x_s varies depending on the subset S.

c. Partial dependence plot

To determine the importance of the target feature in a model, we need to understand both how changing that feature impacts the model's output and the distribution of feature values. PDP provides a visual representation of how specific features, typically one or two at a time, influence the model's predictions while accounting for the remaining features (Friedman 2001; Elith et al. 2008). The partial dependence plot shows the expected response of model output as the feature value of interest is varied, while averaging out the effects of other features in the dataset. These relationships are then graphically depicted to assess the importance of these features. The PDP serves as an interpretable method in machine learning, summarizing the connection between the features of interest and the model's predictions, and it encompasses all instances in the analysis.

Let $f(x)$ be the model's prediction function, where $x = \{x_1, x_2, \ldots, x_N\}$ are the input feature set. The partial dependence ψ_{x_s} for a specific feature x_s is computed by fixing x_s and averaging the model's prediction across all instances in the dataset for the remaining features x_c. For each value of x_s, we calculate the model's prediction while letting the other features vary given by

$$\psi_{x_s} = \frac{1}{n} \sum_{i=1}^{n} f(x_s, x_{ci}), \tag{3}$$

where N is the total number of features and n is the number of observations in the dataset. The $f(x_s, x_{ci})$ is the model's

prediction when the feature x_s is fixed, and the other features x_c take their values from the ith instance in the dataset.

The PDP provides general trends and the overall impact of a feature on the model's predictions, while the SHAP values offer local explanations by quantifying the contribution of each feature to an individual prediction.

4. Classification methods for 3–4 weeks

We applied machine learning techniques, specifically a RF model, to improve global temperature and precipitation forecasts for S2S prediction. Due to the high-dimensional and region-specific nature of climate data, we divided the forecast regions into 23 regions based on weather expert's knowledge. Models for each region were trained and optimized individually, focusing on forecast periods of 3–4 and 5–6 weeks.

In this section, we aim to provide in-depth and interpretable explanations focusing on how individual features influence the predictions in a specific region for temperatures in 3–4 weeks rather than providing detailed explanations for all 23 regions. Specifically, the use of explainable AI tools, such as SHAP values and PDP, allows us to interpret the contribution of individual input features. Note that the same approach could be applied across all regions to derive local interpretations.

a. *RF-based classification models*

RF algorithm (Mitchell 1997; Breiman 2001a; Dietterich 2000; Hill et al. 2020; Louppe 2015) is a machine learning algorithm that combines multiple unrelated decision trees to improve classification accuracy. RF algorithm uses a random resampling technique called "bootstrapping" to generate different subsets of the training data, constructing a separate decision tree for each subset. At each node of a tree, a random subset of features is selected to determine the best split, typically based on criteria like Gini impurity [named after Corrado Gini (1884–1965)] or information gain. The ensemble of decision trees votes on the final classification, with the majority voting being the predicted outcome. This method is robust against overfitting and can handle high-dimensional data efficiently. RF works by aggregating predictions from multiple independent trees, making it more accurate and stable than a single decision tree, which can suffer from high variance. The algorithm's ability to generalize well across different datasets and maintain low bias and low variance makes it particularly useful for complex classification problems.

Algorithm 1 Pseudocode of subseasonal to seasonal prediction algorithms

input: n_{est}, c_{alpha}, $d_{\max}$, c_{func}, λ_{der}, λ_{ext}, λ_{ecm}
  ▷ n_{est}: Number of trees
  c_{alpha}: Complexity parameter used for minimal cost-complexity pruning
  $d_{\max}$: The maximum depth of the tree
  c_{func}: The function to measure the quality of a split
  λ_{der}: Derived features in Table 1, λ_{ext}: External features in Table 2
  λ_{ecm}: ECMWF features in Table 3
for $Y_{\text{vs}} = 2019$ **do** ▷ Y_{vs}: valid start year
  for $i_R = 1$ to N_R **do** ▷ N_R: Number of regions
    Find optimum combination of features $\lambda_{i_R}^{\text{opt}}$ for every region ▷ $\lambda_{i_R}^{\text{opt}} \in \{\lambda_{\text{der}}; \lambda_{\text{ext}}; \lambda_{\text{ecm}}\}$
    R_t: regional train data from train data [region index = i_R, time slice = $(2000, Y_{\text{vs}} - 1)$; $\lambda_{i_R}^{\text{opt}}$]
    R_v: regional validation data from train data [region index = i_R, time slice = $(Y_{\text{vs}}, Y_{\text{vs}} + 1)$; $\lambda_{i_R}^{\text{opt}}$]
    for $i_L = 1$ to N_L **do** ▷ N_L: Number of land
      Extract the latitude and longitude index (i, j) from i_L
      Call a random forest classifier with n_{est}, c_{alpha}, $d_{\max}$, c_{func}
      Build a forest of trees from the training set $[R_t^{(i,j)}, R_v^{(i,j)}]$
      Predict class probabilities for $R_t^{(i,j)}$
      Predict class probabilities for $R_v^{(i,j)}$
      Calculate regional RPSS
    end for
  end for
end for

1) Model structure

Our method begins by dividing the forecast area into 23 regions based on the approximate ranges of similar climates provided by a weather expert, as shown in Fig. 1 and Table 4. In the table, *Europe* and *northern Africa* have two disconnected regions, which span the prime meridian. For each region (1 of 23), lead time (3–4 vs 5–6 weeks), and the response variable to be forecasted (temperature and precipitation), a RF classification model from the Python library scikit-learn (Pedregosa et al. 2011) was trained. Based on domain expert's insights, we undertook to optimize both the feature subset and the fundamental settings (hyperparameters) of the random forest algorithm. Initially, we set the basic parameters of the random forest algorithm to maximize the RPSS. This process involved experimenting with varied splitting criteria, maximum depth, and the number of trees.

The trained models were applied to the test data to predict the tercile class, i.e., whether the test observation is below, at, or above normal. As shown in Fig. 2, one model is trained for each region, each lead-time value (14 and 28 days), and each variable (temperature and precipitation). Further testing showed that using a separate temperature forecasting model for each location works best, while a single precipitation prediction model suffices for all locations for the 28-day lead time. In this case, there would be 90 models in total, as the

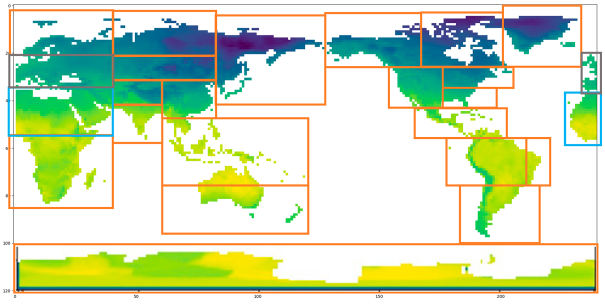

FIG. 1. Domain expert-defined 23 regions for S2S prediction.

TABLE 4. Boundaries of domain expert-defined 23 regions.

Region name	Upper bound	Lower bound	Left bound	Right bound
Scandinavia	0	20	0	40
Europe	20/20	35/35	0/230	40/240
Northern Africa	35/35	57/57	0/225	40/240
Southern Africa	57	90	0	40
North-central Russia	0	20	40	82
Central Russia	20	32	40	82
Himalayas	32	42	40	62
India	42	54	40	62
Eastern China	32	48	62	82
Eastern Russia	0	40	82	127
Indonesia and northern Australia	48	75	62	122
Southern Australia and New Zealand	75	92	62	122
Alaska and western Canada	0	26	127	168
Eastern Canada	0	26	168	203
Greenland	0	26	203	230
Western United States	26	43	151	176
Northeast United States	26	35	176	205
Southeast United States	35	43	176	190
Central America	43	56	163	201
Northern South America	56	76	177	210
Eastern Brazil	56	76	210	220
Southern Central America	76	100	187	210
Antarctica	100	121	0	240

temperature does not have values in the *Antarctic* area. After the regional models are trained, fine-tuned, and cross-validated, we obtain the predicted posterior probabilities of each of the three classes for each location on the test data. We compare the predicted probabilities with climatology and compute the score.

A random forest is composed of nodes, and the nodes seek to optimally split the data to maximize classification accuracy. The classification criterion is the function used to measure the quality of a split, and it allows users to choose between Gini or Entropy reduction (information gain). If a target is a classification outcome taking on values 0, 1, …, $K-1$, for node m, let

$$p_{mk} = \frac{1}{n_m} \sum_{y \in Q_m} I(y = k), \tag{4}$$

where p_{mk} is the predicted posterior probability given the terminal node m, which is the proportion of class k observations at node m. The common measures of impurity are given by

$$\text{Gini-based node selection} : H(Q_m) = \sum_k p_{mk}(1 - p_{mk}), \tag{5}$$

$$\text{Entropy-based node selection} : H(Q_m) = -\sum_k p_{mk} \log(p_{mk}). \tag{6}$$

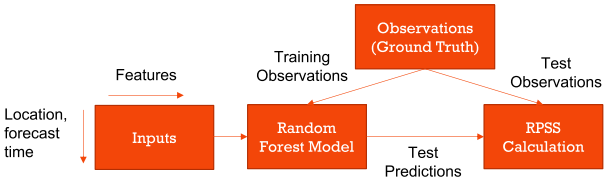

FIG. 2. Flowchart of the work being conducted.

Given that relevant features vary depending on the distinctive characteristics of each region, we demonstrate the hyperparameter tuning process for the *southern Africa* region in this paper. Figure 3 shows the RPSS based on RF classification with varying classification criteria and indicates that the maximum depth and the number of trees tend to have a greater impact than classification criteria. Note that we tested for maximum depths of 1, 2, 5, 10, and 20 and number of trees of 2, 5, 10, 20, 30, and 50. For the *southern Africa* region, we obtain the best regional RPSS with Entropy-criterion-based node selection at a maximum depth = 10 and number of trees = 30.

2) FEATURE IMPORTANCE VIA BACKWARD FEATURE SELECTION

Figure 4 shows how the model performance changes from a reference value when each feature is removed. The reference RPSS is computed using the RF model with the

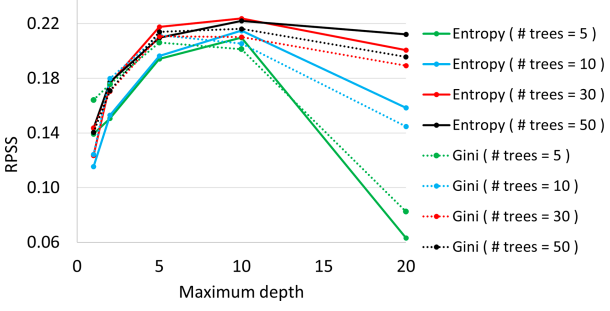

FIG. 3. RPSS of southern Africa for RF with Gini index and Entropy (with varying maximum depths and the number of trees).

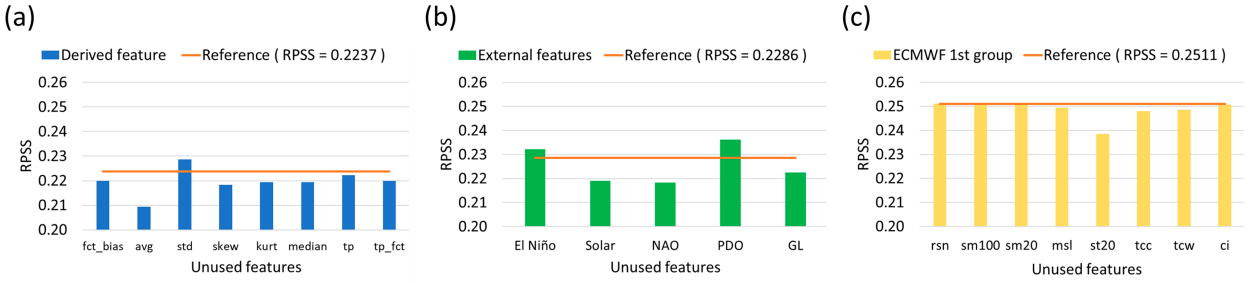

FIG. 4. Useful features with RF (Entropy-criterion-based node selection, maximum depth = 10 and 30 trees) for 3–4 weeks of 2-m temperature in southern Africa. (a) Derived features from S2S observations, (b) features from external observations, and (c) features from ECMWF observations (first group).

Entropy-criterion-based node selection, a maximum tree depth of 10 and number of trees equal to 30 for the 2-m temperature data, for 3–4 weeks. We used the backward feature selection method where we utilize all available features first and then progressively remove the least salient features sequentially to refine and improve the model. The reference RPSS value of 0.2237 in Fig. 4a was generated by using all the derived features, external variables, and the first group of ECMWF forecast features. After finding the best features in the above three feature groups, we investigated the impact of the second group of ECMWF forecast features because the second group has multiple data levels. From the derived features shown in Fig. 4a, it was found that the standard deviation of hindcast realizations when building models was not useful in improving the RPSS because when we removed this feature, the RPSS went up. Other biweekly statistics (biweekly mean, kurtosis, median) and the total precipitation observations, however, were useful predictors. From the features obtained from the external observations in Fig. 4b, we found that the sunspot cycle, NAO, and Great Lake can improve the RPSS. We get an RPSS value of 0.2511 by removing El Niño and PDO, which is used as a reference RPSS in Fig. 4c. From the first group of ECMWF observations, comprised of single-level parameters, as shown in Fig. 4c, the soil temperature in the top 20 cm (st20) is the best performing variable, but besides, all other features except for soil moisture in the top 20 cm (sm20) and sea ice cover (ci) are also helpful in improving the RPSS.

We investigated the second group of features from the ECMWF observations to improve the RPSS, and the results are shown in Fig. 5. Here, the reference value was obtained using the optimal combination of derived features, external observation features, and the first group of ECMWF observations. In this case, we started with the set of features used by the reference and added each of the features at a single air pressure level to determine which air pressure level is best to include. Once we determined the best air pressure level to include, we tried all combinations of these features to determine which gives the best RPSS. Geopotential height (gh), U velocity u, and V velocity v are collected at 10 levels, and we get the highest values at gh = 10 hPa, u = 500 hPa, and v = all levels together for each feature. Features QBO and specific humidity q are collected at seven levels, and the highest values are obtained at QBO = 20 hPa and q = all levels together. Ultimately, the highest RPSS value of 0.2543 was obtained when QBO, gh, u, and q were used together.

3) FEATURE IMPORTANCE AND DEPENDENCY PLOTS VIA SHAP VALUES

SHAP values[2] are used in machine learning to explain how much each feature contributes to a model's prediction. On the vertical axis, the features are ranked from top to bottom in order of importance based on the sum of the magnitudes of their SHAP values across all data points, as shown in Fig. 6a. In other words, the higher a feature appears on the plot, the greater its influence on the model's predictions. In Fig. 6b, both the significance of features and their impact on predictions are illustrated. Each data point on the summary plot represents a SHAP value for a specific feature within an instance. Along the horizontal axis, the distribution of each feature's impact on the model's predictions is displayed. The color of the dots corresponds to original feature values, with red indicating high values and blue indicating low values.

For example, in features such as the soil temperature in the top 20 cm (st20) and QBO at 20 hPa (QBO_20), the distribution of SHAP values is such that the red dots are to the left of the origin and the blue dots are to the right of the origin in the plot. This means that lowering these feature values contributes to an increase in the predicted model output, while increasing the feature values tends to reduce it. On the other hand, the distribution of SHAP values, such as the sunspot cycles (Solar) and total cloud cover (tcc), where the red dots are to the right of the origin and the blue dots are to the left

[2] We compute the approximation of SHAP values as implemented in Scott Lundberg's Python SHAP library (Lundberg and Lee 2017), using version 0.43.0. Specifically, we employ TreeExplainer, which is optimized for tree-based models. This method approximates SHAP values efficiently by leveraging the tree model structure, significantly reducing the computation time when compared to computing the exact SHAP values on general models. The computation time is approximately 35 minutes for observations with 193 185 rows × 49 columns for the *southern Africa* region, and the computation time depends on the dataset size and model complexity.

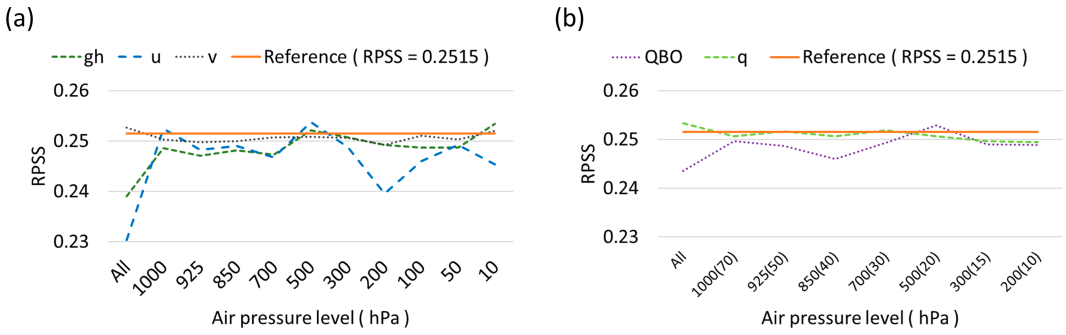

FIG. 5. Useful features of ECMWF observations (second group) for 3–4 weeks of 2-m temperature in southern Africa. (a) Features with 10 pressure levels for gh, u, and v and (b) features with 7 pressure levels for QBO and q.

of the origin, suggests that increasing the feature values increases the predicted model output. In other words, red dots positioned to the right indicate that higher feature values are associated with positive SHAP contributions to the model output, whereas red dots on the left indicate negative contributions. These SHAP patterns represent localized, model-derived attributes rather than global statistical correlations and can be used to identify predictive signals learned by the model.

Note that SHAP values are not gridded within regions. The same model is applied across all points within a region, resulting in a single SHAP value for the region. This means that the SHAP value represents the average importance of each feature across the region, rather than varying spatially. Although this method does not account for potential local variability in temperature and precipitation, it simplifies the model and provides a consistent metric for a region.

The PDP is a tool to visualize the relationship between the features and the predicted outcomes in complex models. The PDP provides how a specific feature affects the model's predictions on average, making it a crucial method for model interpretation and explanation. Figure 7 shows the partial dependence plots of three most important features (st20, Solar, and QBO = 20 hPa) and three least important features (rsn, krt, and skw). For example, Fig. 7a shows the distribution of soil temperature in the top 20 cm (st20) and the expected value of the model when applied to the dataset along with the horizontal axis. The vertical axis represents how the average value of the prediction changes as st20 is varied. The dependence plot suggests a monotonic S-shaped relationship between st20 and the expected model prediction. Note that the partial dependence plot is the average value of the model output when we fix st20 to a given value. The three least important features have a relatively narrower range of partial dependence values than the three important features.

Table 5 shows the regional RPSS for 2-m temperature on the 3–4-week dataset. The reference RPSS of 0.2237 is based on the RF model with the Entropy-criterion-based node selection, a maximum tree depth of 10 and number of trees equal to 30 using all the derived features, external variables, and the first group of ECMWF forecast features. We obtained an RPSS improvement to 0.2286 with the optimal derived features, and we further improved the RPSS value to 0.2511 by using the optimal external features. In the ECMWF features, snow density (rsn), soil moisture in the top 100 cm (sm100), mean sea level pressure (msl), soil temperature in the top 20 cm (st20), total cloud cover (tcc), and total column water (tcw) are useful for improving the RPSS, and we get 0.2515 by removing sm20. The highest score for 2-m temperature for the 3–4-week dataset, 0.2543, was obtained when QBO, geopotential height (gh), U velocity u, and specific humidity q were used together. This is the best improvement in the RPSS value found so far on this dataset, and we improved the RPSS by 13.7% from 0.2237 in *southern Africa* through this tuning process.

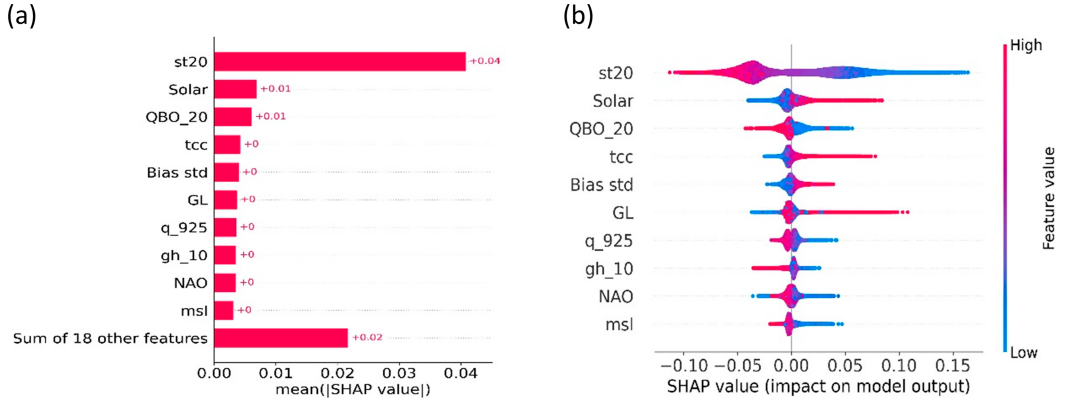

FIG. 6. SHAP values for 3–4 weeks of 2-m temperature in southern Africa (top 10 important features). (a) SHAP mean values and (b) impact on model output.

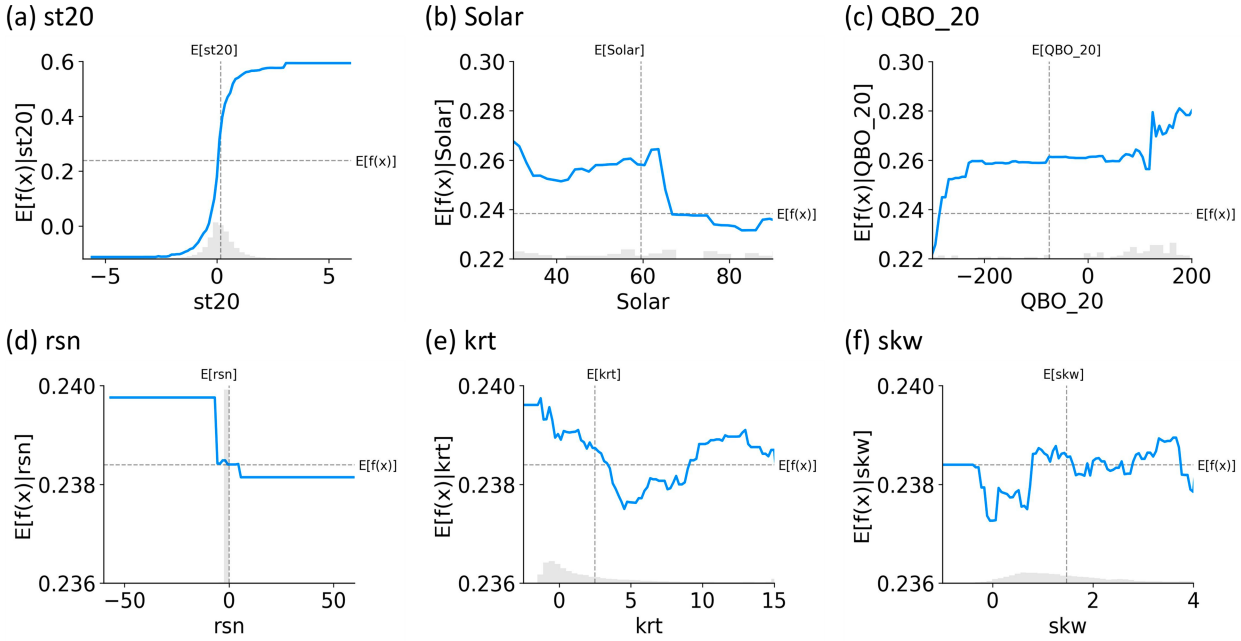

FIG. 7. PDPs of 2-m temperature for 3–4 weeks in southern Africa. (a)–(c) Three most important features from left to right; (d)–(f) three least important features from left to right.

Figure 8 shows the confusion matrices for all predictions in three regions (*eastern Brazil*, *Central America*, *southern Africa*) with the highest RPSS and three regions (*Southeast United States*, *southern Central America*, *western United States*) with the lowest RPSS. These confusion matrices indicate that the high category is consistently well predicted across both high and low RPSS regions, showing that the model performs best at predicting high category. However, the model performs poorly at identifying the medium category, as evidenced by the low probabilities for this class in all regions. This suggests that the model may not adequately differentiate between low and medium categories, often defaulting to either high or low or that WMO challenge designers could have removed the medium category altogether. These results show that the model very rarely predicts the medium category as the most likely, but the RPSS results show that the predictions are significantly better than climatography.

Figure 9 and Table 6 show the RPSS in each region, and Table 6 also provides both the hyperparameters used for the random forest model for each region and the detailed pressure levels for the second group of ECMWF observations. Note that the *Antarctica* region is not considered in 2-m temperature dataset. The overall RPSS of 2-m temperature for 3–4 weeks is 0.123. We also compared our results to the ECMWF prediction of 2-m temperature. The ECMWF forecasts are real valued temperatures with multiple realizations. To convert these to categorical variables in line with our machine learning predictions, we categorized the temperatures into low, medium, and high in the same way that we created our machine learning labels. We then averaged the values across the realizations. This allows us to calculate the RPSS for the ECMWF forecasts as a comparison to our methods, as shown in Table 6. The only region where ECMWF forecast had a higher RPSS was *north-central Russia*, and the ECMWF RPSS was only slightly higher. In several other regions, such as *southern Africa*, and *Central America*, the machine learning forecasts significantly outperform ECMWF, which is performing much worse than using climatographic historical averages.

TABLE 5. Regional RPSS of 2-m temperature for 3–4 weeks in southern Africa.

Name	Features	RPSS
Reference with RF model with Entropy-criterion-based node selection, a maximum tree depth of 10 and 30 trees	All derived features, external variables, and the first group of ECMWF forecast	0.2237
Reference with optimal derived features	$N_d = 9$, $N_y = 10$, bias statistics, biweekly statistics (avg, skw, krt, med)	0.2286
Reference with derived features and external features	Solar, GL, NAO	0.2511
Reference with derived features, external features, and ECMWF features	rsn, sm100, msl, st20, tcc, tcw, QBO, gh, u, q	0.2543

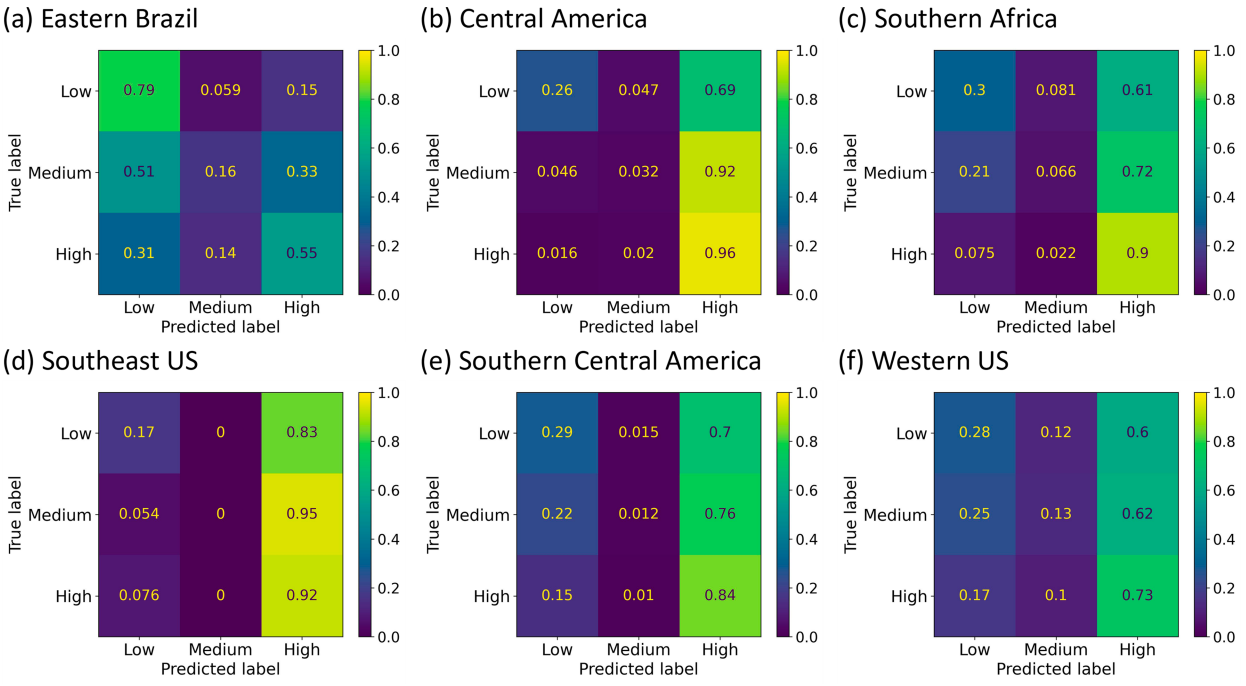

FIG. 8. Confusion matrices of 2-m temperature for 3–4 weeks. (a)–(c) Three regions with the highest RPSS from left to right; (d)–(f) three regions with the lowest RPSS from left to right.

This shows that although the confusion matrices in Fig. 8 indicate that the performance of the machine learning model is somewhat underwhelming, it is still significantly better than the state-of-the-art S2S forecasting methods. This highlights the challenge of S2S forecasting, i.e., the chaotic nature of atmospheric dynamics, and shows that machine learning methods can be a step toward overcoming this challenge.

The SHAP values for all regions and all features are shown in Fig. 10. For predicting 2-m temperature for 3–4 weeks, we used 21 features ($N_d = 9$, $N_y = 10$, scaled values of hindcast mean and standard deviation) as default features and up to 90 features can be applied depending on the region. The features that are not used are left blank. Although regional variations exist, the soil temperature in the top 20 cm (st20), the total column water (tcw), the sunspot cycles (Solar), El Niño, and the PDO generally show high SHAP values. Specifically, st20 is the most influential global factor affecting our model predictions across various regions, likely due to its critical role in land–atmosphere interactions. For instance, the soil temperature significantly impacts energy exchange processes, including heat flux and evapotranspiration, which are essential for both local and global climate systems.

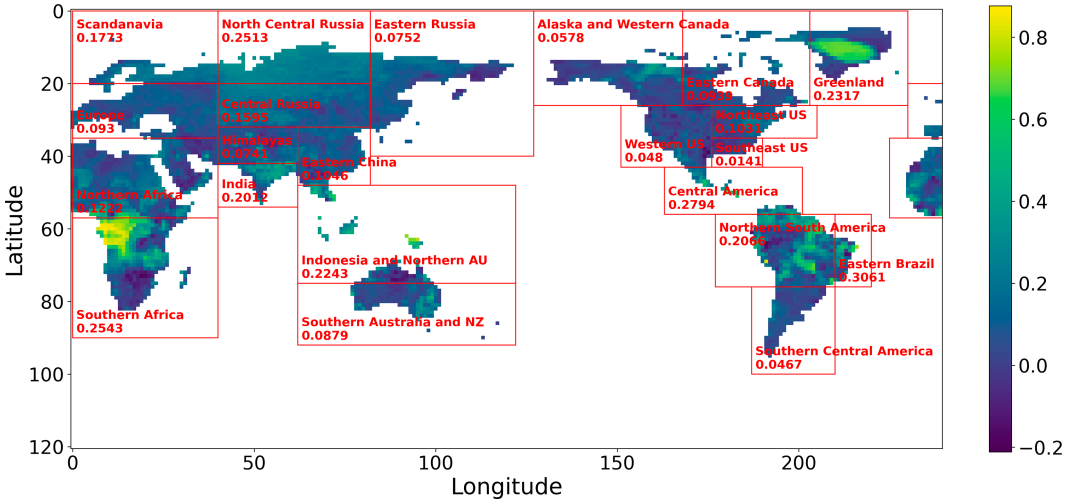

FIG. 9. Overall RPSS of 2-m temperature for 3–4 weeks.

TABLE 6. Parameters of RF classifier and detailed pressure levels for 2-m temperature for 3–4 weeks.

	RF classifier			ECMWF observations (second group)					Regional RPSS	
Region	No. of trees	Max depth	Criterion	QBO	gh	*u*	*v*	*q*	ML-based	ECMWF-based
Scandinavia	50	10	Gini					850	0.1773	0.0752
Europe	20	10	Entropy		All				0.0930	−0.0240
Northern Africa	50	10	Entropy		700		All		0.1222	−0.6072
Southern Africa	30	10	Entropy	20	10	500		All	0.2543	−0.8629
North-central Russia	20	5	Gini	70	925		10	All	0.2513	0.2588
Central Russia	5	5	Gini		All	700		All	0.1595	−0.0196
Himalayas	50	10	Gini		500		1000	700	0.0741	−0.6250
India	50	20	Entropy	70	500		50		0.2012	−0.2922
Eastern China	50	10	Entropy					All	0.1046	−0.4137
Eastern Russia	50	5	Entropy		All	700	925	All	0.0752	−0.1517
Indonesia and northern Australia	20	10	Entropy	40	All				0.2243	−0.8304
Southern Australia and New Zealand	10	5	Entropy			850			0.0879	−0.0636
Alaska and western Canada	30	10	Entropy		50				0.0578	−0.0880
Eastern Canada	5	5	Entropy		850	850			0.0939	−0.0859
Greenland	20	5	Entropy			50			0.2317	0.0848
Western United States	5	10	Gini					All	0.0480	−0.3364
Northeast United States	30	10	Entropy		All			850	0.1031	−0.0947
Southeast United States	20	2	Gini					1000	0.0141	−0.0577
Central America	50	10	Entropy		All				0.2794	−1.3421
Northern South America	10	5	Entropy		700	925	100		0.2066	−1.0778
Eastern Brazil	30	10	Entropy						0.3061	−0.5544
Southern Central America	10	5	Gini	20	300			925	0.0467	−0.3888

b. *Clustering methods versus domain-expert-recommended region definition*

So far, we have shown the results using regions defined by a subject matter expert. In this section, we investigate machine learning clustering methods to determine the forecast regions and show that they underperform compared to the expert-defined regions. The *k*-means algorithm groups data into *n* clusters by seeking to separate samples into *n* groups of nearly equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares. This algorithm requires the number of clusters, *n* to be specified. It scales well to large numbers of samples and has been used across a wide range of application areas in many different fields.

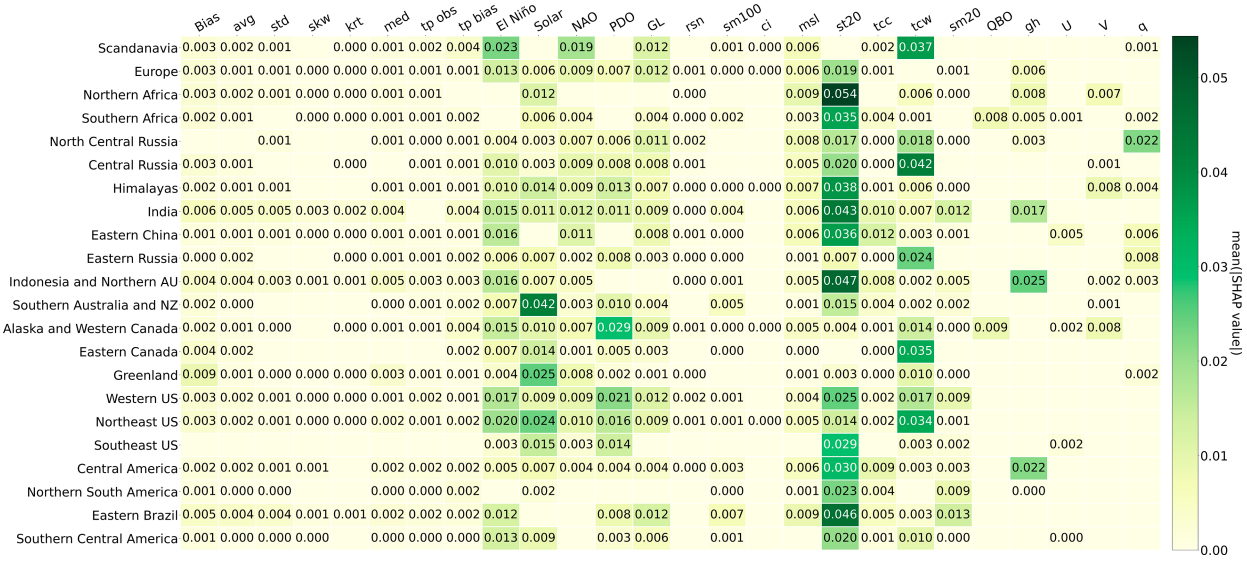

FIG. 10. Useful features of 2-m temperature for 3–4 weeks based on the SHAP map.

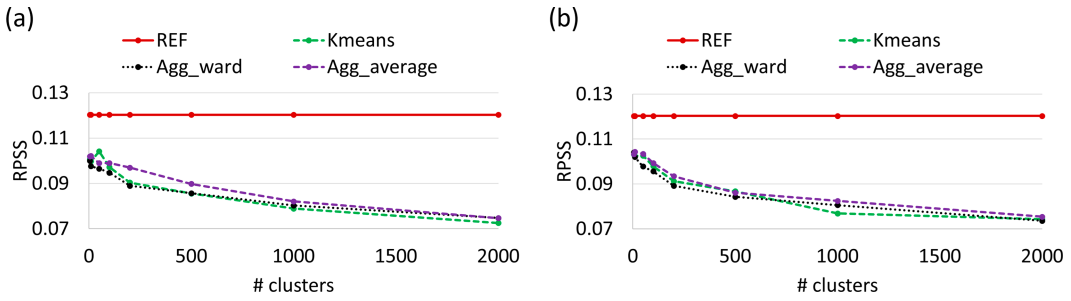

FIG. 11. RPSS values based on (a) residual clustering method and (b) observation clustering method.

Hierarchical clustering is a general family of clustering algorithms that build nested clusters by merging or splitting them successively. This hierarchy of clusters is represented as a tree, called a dendrogram. The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample. Ward minimizes the sum of squared differences within all clusters. It is a variance-minimizing approach and in this sense is similar to the *k*-means objective function but tackled with an agglomerative hierarchical approach. Agglomerative clustering can also scale to a large number of samples when it is used jointly with a connectivity matrix, but is computationally expensive when no connectivity constraints are added between samples.

Figure 11 shows the RPSS values of clustering methods. All clustering methods show worse performance compared to the reference (REF), suggesting that domain expertise-based region definition is better than ML-based region definition in this application.

c. *Comparison with other ML methods*

In this section, we describe the machine learning methods other than random forest that we experimented with. The multilayer perceptron (MLP) is a feedforward artificial neural network model that maps input datasets to a set of appropriate outputs. An MLP consists of multiple layers, and each layer is fully connected to the following one. The nodes of the layers are neurons with nonlinear activation functions, except for the nodes of the input layer. Between the input and the output layer, there may be one or more nonlinear hidden layers. The k-nearest neighbors (kNNs) algorithm is a nearest neighbor classification model in which we can alter both the distance metric and the number of nearest neighbors. The choice of the value of *k* is dependent on the data. The gradient boosting (GB) algorithm builds a prediction model in the form of an ensemble of weak prediction models. The aim of the GB procedure is to minimize the bias error of the model. Table 7 shows the hyperparameters used for the four machine learning methods, as well as their performance on training, validation, and test sets.

5. Discussion of 3–4-week and 5–6-week results

As an extension of our work in the 2021 WMO S2S AI challenge, we present hybrid machine learning algorithms, integrating domain expert-defined geographic regions with explainable AI techniques, such as SHAP values and PDP. This approach enhances prediction accuracy and interpretability by incorporating the expert knowledge of climate dynamics, distinguishing our method from traditional data-driven clustering techniques to define the regions.

Specifically, we show that the model's reliance on the soil temperature in the top 20 cm (st20) across most regions highlights the importance of st20 as a predictor in forecasting 2-m temperature for 3–4 weeks. However, this could also imply that the model is overrelying on st20, potentially due to the insufficient inclusion of other relevant features, such as humidity, cloud cover, or vegetation indices, in the training data. To ensure that the model is appropriately leveraging st20, a deeper evaluation is needed to verify whether its influence is balanced and accurately weighted.

TABLE 7. Comparison with other ML methods for 2-m temperature for 3–4 weeks in southern Africa.

Parameter	MLP	kNN	GB	RF
Metric	logistic	minkowski	friedman_mse	Entropy
Random state	1	1	0	None
No. of estimators	None	50	20	30
Maximum depth	None	None	5	10
Initial learning rate	0.001	None	0.1	None
RPSS for training (computation time)	0.220 (306 s)	0.151 (100 s)	0.122 (348 s)	0.184 (145 s)
RPSS for validation	0.284	0.144	0.329	0.345
RPSS for testing (computation time)	0.128 (1.5 s)	0.068 (88.8 s)	0.243 (1.3 s)	0.254 (2.6 s)

Table 8 shows the S2S prediction results based on temperature and precipitation for periods of 3–4 and 5–6 weeks, and comparison of RPSS results with the ML model of CRIMS2S team submitted to WMO S2S artificial intelligence challenge. Our proposed method shows the improved performance from 23.9% to up to 40% compared to the model that performed best in WMO challenge, but our method shows very poor performance for total precipitation for 5–6 weeks.

When compared to the total precipitation prediction, the 2-m temperature is easier to predict due to its spatial and temporal stability and its strong correlation with atmospheric variables, like surface temperature and high-resolution, globally consistent data. These factors enable effective model training and generalization, making the 2-m temperature a relatively stable and reliable response variable to predict, even for an extended forecast period of 5–6 weeks.

The poor performance of total precipitation predictions for 5–6 weeks can be attributed to several factors. First, precipitation is inherently more variable and nonlinear than a variable such as 2-m temperature. Its prediction depends on complex atmospheric and oceanic interactions, which become increasingly uncertain over extended forecast periods. Additionally, the training dataset may lack sufficient representation of precipitation patterns that are specific to the 5–6-week period, leading to suboptimal generalization by the ML model. Another challenge is the relatively weak correlation between precipitation and the primary predictors used in the model, such as temperature, when compared to other variables like humidity or cloud cover, which may not have been fully integrated.

In this paper, we use SHAP values to investigate the individual contribution of each feature in S2S predictions, but SHAP values are not gridded within regions. For future work, we plan to explore the application of gridded SHAP values to better capture detailed spatial variations within regions exhibiting more complex weather patterns. This approach can provide a more detailed understanding of local variability, enhancing the model's accuracy in representing spatial differences across diverse climatic conditions.

As another limitation in our current approach, we recognize that some predictive signals are stronger in specific seasons. We will aim to address this limitation by using SHAP values to analyze seasonal variations. We intend to focus on specific geographical regions with high ML accuracy, defined by domain experts, particularly those influenced by climate oscillations, such as ENSO, PDO, NAO, and the Madden–Julian oscillation (MJO). It will help us enhance regional accuracy and identify key patterns between model accuracy and feature importance, ultimately improving the trustworthiness and robustness of our ML models. Given the significant role of the MJO, especially in modulating tropical convection and influencing the global weather systems, incorporating features, such as its phase and amplitude, may enhance predictive accuracy in S2S forecasts.

6. Conclusions

In this paper, we presented machine learning (ML) techniques to enhance the accuracy of global temperature and precipitation forecasts by combining predictions from the numerical weather models with historical data. Our work focused on forecast periods of 3–4 and 5–6 weeks, which are notably challenging for subseasonal to seasonal (S2S) prediction. Specifically, the application of a random forest (RF) classification model to 23 regions resulted in a significant improvement, with a 12.3% improvement in ranked probability skill scores (RPSSs) for temperature predictions and a 4.2% improvement for precipitation predictions over the 3–4-week period. Although the improvements were more moderate for the 5–6-week forecast period, with a 5.7% increase in temperature prediction accuracy and a marginal 0.1% improvement in precipitation accuracy, these results represent important advancements in addressing the predictability challenges for these critical time scales.

In this paper, explainable AI tools, such as Shapley additive explanation (SHAP) values and partial dependence plot (PDP), provide a new level of transparency and interpretability to our forecasts, allowing us to better understand which climate variables are the most influential for each region. However, trustworthiness in predictive models is a complex and multifaceted concept that goes beyond explainability. It includes aspects such as robustness, fairness, reliability, and generalizability. Additional methodologies and further research are necessary to comprehensively evaluate and ensure trustworthiness across these dimensions.

Given the challenges inherent in S2S forecasts, these moderate improvements could be used to improve long-term decision-making in areas, such as agriculture and water resource management, that depend heavily on the temperature and precipitation levels. Enhanced forecast precision could inform better resource allocation, risk mitigation strategies, and more effective planning in areas prone to climatic variability, such as optimizing irrigation schedules or preparing for potential droughts.

To further build on these findings, several avenues for future research are suggested. First, hybrid models combining multiple machine learning algorithms with numerical weather prediction models could be explored to improve forecast

TABLE 8. Comparison with WMO models.

	Global RPSS (2-m temperature)		Global RPSS (total precipitation)	
Lead time (weeks)	WMO-CRIMS2S	UConn team	WMO-CRIMS2S	UConn team
3–4	0.090	0.123	0.030	0.042
5–6	0.046	0.057	0.017	0.001

accuracy. In particular, integrating deep learning techniques with RF models may offer enhanced capabilities for capturing the nonlinearities and complex relationships within high-dimensional climate data. Second, there is a need to incorporate uncertainty quantification into S2S models to provide decision-makers with confidence intervals and risk assessments, thus enhancing the robustness of predictions. Finally, given the evolving nature of climate patterns due to global warming, future research should explore how machine learning models can adapt to these changes. One promising approach is the use of transfer learning, which could allow models to generalize across different regions and adjust to shifting climatic conditions.

Acknowledgments. This work was supported, in part, by the U.S. Office of Naval Research and Naval Research Laboratory under Grants N00014-18-1-1238, N00014-21-1-2187, and N00173-22-1-G005. We would like to acknowledge contributions from Peter Willett from the University of Connecticut, and Paolo Braca and Leonardo Millefiori from the NATO STO CMRE.

Data availability statement. ECMWF forecast datasets are openly available from ECMWF Confluence (https://confluence.ecmwf.int/display/S2S/Parameters). Data and ML model algorithm used in this study can be found in Zenodo (https://doi.org/10.5281/zenodo.10275034).

REFERENCES

Arakawa, A., 2004: The cumulus parameterization problem: Past, present, and future. *J. Climate*, **17**, 2493–2525, https://doi.org/10.1175/1520-0442(2004)017<2493:RATCPP>2.0.CO;2.

Baldwin, M. P., and Coauthors, 2001: The quasi-biennial oscillation. *Rev. Geophys.*, **39**, 179–229, https://doi.org/10.1029/1999RG000073.

Barnston, A. G., and R. E. Livezey, 1987: Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Mon. Wea. Rev.*, **115**, 1083–1126, https://doi.org/10.1175/1520-0493(1987)115<1083:CSAPOL>2.0.CO;2.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, https://doi.org/10.1038/nature14956.

Belayneh, A., J. Adamowski, B. Khalil, and B. Ozga-Zielinski, 2014: Long-term SPI drought forecasting in the Awash River basin in Ethiopia using wavelet neural network and wavelet support vector regression models. *J. Hydrol.*, **508**, 418–429, https://doi.org/10.1016/j.jhydrol.2013.10.052.

Bender, M. A., and I. Ginis, 2000: Real-case simulations of hurricane–ocean interaction using a high-resolution coupled model: Effects on hurricane intensity. *Mon. Wea. Rev.*, **128**, 917–946, https://doi.org/10.1175/1520-0493(2000)128<0917:RCSOHO>2.0.CO;2.

Bi, K., L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, 2022: Pangu-weather: A 3D high-resolution model for fast and accurate global weather forecast. arXiv, 2211.02556v1, https://doi.org/10.48550/arXiv.2211.02556.

Bishop, C. M., 2006: *Pattern Recognition and Machine Learning.* Vol. 4, Springer, 738 pp.

Breiman, L., 2001a: Random forests. *Mach. Learn.*, **45**, 5–32, https://doi.org/10.1023/A:1010933404324.

——, 2001b: Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Stat. Sci.*, **16**, 199–231, https://doi.org/10.1214/ss/1009213726.

Buizza, R., 2002: Chaos and weather prediction January 2000. European Centre for Medium-Range Weather Meteorological Training Course Lecture Series ECMWF, 28 pp., https://msi.ttu.ee/~juri.elken/Predicting_and_Chaos.pdf.

Bzdok, D., N. Altman, and M. Krzywinski, 2018: Statistics versus machine learning. *Nat. Methods*, **15**, 233–234, https://doi.org/10.1038/nmeth.4642.

Chai, T., and R. R. Draxler, 2014: Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.*, **7**, 1247–1250, https://doi.org/10.5194/gmd-7-1247-2014.

Chen, L., B. Han, X. Wang, J. Zhao, W. Yang, and Z. Yang, 2023: Machine learning methods in weather and climate applications: A survey. *Appl. Sci.*, **13**, 12019, https://doi.org/10.3390/app132112019.

Chen, T., and C. Guestrin, 2016: Xgboost: A scalable tree boosting system. *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, New York, NY, Association for Computing Machinery, 785–794, https://doi.org/10.1145/2939672.2939785.

Christoph, M., 2020: *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* Leanpub, 320 pp.

Chrit, M., and M. Majdi, 2024: Operational wind and turbulence nowcasting capability for advanced air mobility. *Neural Comput. Appl.*, **36**, 10 637–10 654, https://doi.org/10.1007/s00521-024-09614-0.

Cohen, J., D. Coumou, J. Hwang, L. Mackey, P. Orenstein, S. Totz, and E. Tziperman, 2019: S2S reboot: An argument for greater inclusion of machine learning in subseasonal to seasonal forecasts. *Wiley Interdiscip. Rev.: Climate Change*, **10**, e00567, https://doi.org/10.1002/wcc.567.

Dai, A., and T. M. L. Wigley, 2000: Global patterns of ENSO-induced precipitation. *Geophys. Res. Lett.*, **27**, 1283–1286, https://doi.org/10.1029/1999GL011140.

Dawson, A., and T. N. Palmer, 2015: Simulating weather regimes: Impact of model resolution and stochastic parameterization. *Climate Dyn.*, **44**, 2177–2193, https://doi.org/10.1007/s00382-014-2238-x.

de Oliveira e Lucas, P., M. A. Alves, P. C. de Lima e Silva, and F. G. Guimarães, 2020: Reference evapotranspiration time series forecasting with ensemble of convolutional neural networks. *Comput. Electron. Agric.*, **177**, 105700, https://doi.org/10.1016/j.compag.2020.105700.

Dietterich, T. G., 2000: An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Mach. Learn.*, **40**, 139–157, https://doi.org/10.1023/A:1007607513941.

Dikshit, A., B. Pradhan, and A. M. Alamri, 2020: Short-term spatio-temporal drought forecasting using random forests model at New South Wales, Australia. *Appl. Sci.*, **10**, 4254, https://doi.org/10.3390/app10124254.

Doblas-Reyes, F. J., J. García-Serrano, F. Lienert, A. P. Biescas, and L. R. L. Rodrigues, 2013: Seasonal climate predictability and forecasting: Status and prospects. *Wiley Interdiscip. Rev.: Climate Change*, **4**, 245–268, https://doi.org/10.1002/wcc.217.

Ehrendorfer, M., 1997: Vorhersage der Unsicherheit numerischer Wetterprognosen: Eine Übersicht. *Meteor. Z.*, **6**, 147–183, https://doi.org/10.1127/metz/6/1997/147.

Elith, J., J. R. Leathwick, and T. Hastie, 2008: A working guide to boosted regression trees. *J. Anim. Ecol.*, **77**, 802–813, https://doi.org/10.1111/j.1365-2656.2008.01390.x.

Epstein, E. S., 1969a: 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.

——, 1969b: Stochastic dynamic prediction. *Tellus*, **21**, 739–759, https://doi.org/10.3402/tellusa.v21i6.10143.

Friedman, J. H., 2001: Greedy function approximation: A gradient boosting machine. *Ann. Stat.*, **29**, 1189–1232, https://doi.org/10.1214/aos/1013203451.

Gneiting, T., A. E. Raftery, A. H. Westveld III, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, https://doi.org/10.1175/MWR2904.1.

Goodfellow, I., Y. Bengio, and A. Courville, 2016: *Deep Learning.* MIT Press, 775 pp.

Han, J.-Y., S.-W. Kim, C.-H. Park, and S.-W. Son, 2023: Ensemble size versus bias correction effects in subseasonal-to-seasonal (S2S) forecasts. *Geosci. Lett.*, **10**, 37, https://doi.org/10.1186/s40562-023-00292-9.

Hao, Z., V. P. Singh, and Y. Xia, 2018: Seasonal drought prediction: Advances, challenges, and future prospects. *Rev. Geophys.*, **56**, 108–141, https://doi.org/10.1002/2016RG000549.

Hill, A. J., G. R. Herman, and R. S. Schumacher, 2020: Forecasting severe weather with random forests. *Mon. Wea. Rev.*, **148**, 2135–2161, https://doi.org/10.1175/MWR-D-19-0344.1.

Hudson, D., O. Alves, H. H. Hendon, and A. G. Marshall, 2011: Bridging the gap between weather and seasonal forecasting: Intraseasonal forecasting for Australia. *Quart. J. Roy. Meteor. Soc.*, **137**, 673–689, https://doi.org/10.1002/qj.769.

Jordan, M. I., and T. M. Mitchell, 2015: Machine learning: Trends, perspectives, and prospects. *Science*, **349**, 255–260, https://doi.org/10.1126/science.aaa8415.

Kashinath, K., and Coauthors, 2021: Physics-informed machine learning: Case studies for weather and climate modelling. *Philos. Trans. Roy. Soc.*, **A379**, 20200093, https://doi.org/10.1098/rsta.2020.0093.

Koster, R. D., and Coauthors, 2010: Contribution of land surface initialization to subseasonal forecast skill: First results from a multi-model experiment. *Geophys. Res. Lett.*, **37**, L02402, https://doi.org/10.1029/2009GL041677.

Kurth, T., and Coauthors, 2023: FourCastNet: Accelerating global high-resolution weather forecasting using adaptive Fourier neural operators. *Proc. Platform for Advanced Scientific Computing Conf.*, New York, NY, Association for Computing Machinery, 1–11, https://doi.org/10.1145/3592979.3593412.

Lam, R., and Coauthors, 2023: Graphcast: Learning skillful medium-range global weather forecasting. arXiv, 2212.12794v2, https://doi.org/10.48550/arXiv.2212.12794.

Lee, Y.-G., J.-Y. Oh, D. Kim, and G. Kim, 2023: SHAP value-based feature importance analysis for short-term load forecasting. *J. Electr. Eng. Technol.*, **18**, 579–588, https://doi.org/10.1007/s42835-022-01161-9.

Li, W., J. Chen, L. Li, H. Chen, B. Liu, C.-Y. Xu, and X. Li, 2019: Evaluation and bias correction of S2S precipitation for hydrological extremes. *J. Hydrometeor.*, **20**, 1887–1906, https://doi.org/10.1175/JHM-D-19-0042.1.

Lorenc, A. C., 1986: Analysis methods for numerical weather prediction. *Quart. J. Roy. Meteor. Soc.*, **112**, 1177–1194, https://doi.org/10.1256/smsqj.47413.

Louppe, G., 2015: Understanding random forests: From theory to practice. arXiv, 1407.7502v3, https://doi.org/10.48550/arXiv.1407.7502.

Lundberg, S. M., and S.-I. Lee, 2017: A unified approach to interpreting model predictions. *Proc. 31st Int. Conf. on Neural Information Processing Systems*, Red Hook, NY, Curran Associates Inc., 4768–4777, https://doi.org/10.5555/3295222.3295230.

——, and Coauthors, 2020: From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.*, **2**, 56–67, https://doi.org/10.1038/s42256-019-0138-9.

Mantua, N. J., and S. R. Hare, 2002: The Pacific decadal oscillation. *J. Oceanogr.*, **58**, 35–44, https://doi.org/10.1023/A:1015820616384.

Marcílio, W. E., and D. M. Eler, 2020: From explanations to feature selection: Assessing SHAP values as feature selection mechanism. *33rd SIBGRAPI Conf. on Graphics, Patterns and Images (SIBGRAPI)*, Porto de Galinhas, Brazil, Institute of Electrical and Electronics Engineers, 340–347, https://doi.org/10.1109/SIBGRAPI51738.2020.00053.

Meehl, G. A., J. M. Arblaster, K. Matthes, F. Sassi, and H. van Loon, 2009: Amplifying the Pacific climate system response to a small 11-year solar cycle forcing. *Science*, **325**, 1114–1118, https://doi.org/10.1126/science.1172872.

Mendoza, P. A., B. Rajagopalan, M. P. Clark, K. Ikeda, and R. M. Rasmussen, 2015: Statistical postprocessing of high-resolution regional climate model output. *Mon. Wea. Rev.*, **143**, 1533–1553, https://doi.org/10.1175/MWR-D-14-00159.1.

Mitchell, T. M., 1997: *Machine Learning* (1st. ed.). McGraw-Hill, 432 pp.

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.

Navon, I. M., 2009: Data assimilation for numerical weather prediction: A review. *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications*, S. K. Park and L. Xu, Eds., Springer, 21–65, https://doi.org/10.1007/978-3-540-71056-1_2.

Palmer, T. N., and D. L. T. Anderson, 1994: The prospects for seasonal forecasting—A review paper. *Quart. J. Roy. Meteor. Soc.*, **120**, 755–793, https://doi.org/10.1002/qj.49712051802.

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.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Price, I., and Coauthors, 2023: Gencast: Diffusion-based ensemble forecasting for medium-range weather. arXiv, 2312.15796v2, https://doi.org/10.48550/arXiv.2312.15796.

Rasp, S., and Coauthors, 2024: WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *J. Adv. Model. Earth Syst.*, **16**, e2023MS004019, https://doi.org/10.1029/2023MS004019.

Robertson, A. W., A. Kumar, M. Peña, and F. Vitart, 2015: Improving and promoting subseasonal to seasonal prediction. *Bull. Amer. Meteor. Soc.*, **96**, ES49–ES53, https://doi.org/10.1175/BAMS-D-14-00139.1.

Ropelewski, C. F., and M. S. Halpert, 1986: North American precipitation and temperature patterns associated with the El Niño/Southern Oscillation (ENSO). *Mon. Wea. Rev.*, **114**,

2352–2362, https://doi.org/10.1175/1520-0493(1986)114<2352:NAPATP>2.0.CO;2.

Schmidhuber, J., 2015: Deep learning in neural networks: An overview. *Neural Networks*, **61**, 85–117, https://doi.org/10.1016/j.neunet.2014.09.003.

Shalev-Shwartz, S., and S. Ben-David, 2014: *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 397 pp.

Shmueli, G., 2010: To explain or to predict? *Statist. Sci.*, **25**, 289–310, https://doi.org/10.1214/10-STS330.

Slingo, J., and T. Palmer, 2011: Uncertainty in weather and climate prediction. *Philos. Trans. Roy. Soc.*, **A369**, 4751–4767, https://doi.org/10.1098/rsta.2011.0161.

Staniak, M., and P. Biecek, 2019: Explanations of model predictions with live and breakDown Packages. *R J.*, **10**, 395, https://doi.org/10.32614/RJ-2018-072.

Thompson, P. D., 1957: Uncertainty of initial state as a factor in the predictability of large scale atmospheric flow patterns. *Tellus*, **9**, 275–295, https://doi.org/10.1111/j.2153-3490.1957.tb01885.x.

Verbitski, A., and Coauthors, 2017: Amazon aurora: Design considerations for high through-put cloud-native relational databases. *Proc. 2017 ACM Int. Conf. on Management of Data*, New York, NY, Association for Computing Machinery, 1041–1052, https://doi.org/10.1145/3035918.3056101.

Vitart, F., A. W. Robertson, and D. L. T. Anderson, 2012: Subseasonal to seasonal prediction project: Bridging the gap between weather and climate. *WMO Bull.*, **61**, 23.

——, 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.

Wang, H., Q. Liang, J. T. Hancock, and T. M. Khoshgoftaar, 2024: Feature selection strategies: A comparative analysis of SHAP-value and importance-based methods. *J. Big Data*, **11**, 44, https://doi.org/10.1186/s40537-024-00905-w.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2007: The discrete brier and ranked probability skill scores. *Mon. Wea. Rev.*, **135**, 118–124, https://doi.org/10.1175/MWR3280.1.

Weyn, J. A., D. R. Durran, R. Caruana, and N. Cresswell-Clay, 2021: Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. *J. Adv. Model. Earth Syst.*, **13**, e2021MS002502, https://doi.org/10.1029/2021MS002502.

White, C. J., S. W. Franks, and D. McEvoy, 2015: Using subseasonal-to-seasonal (S2S) extreme rainfall forecasts for extended-range flood prediction in Australia. *Proc. Int. Assoc. Hydrol. Sci.*, **370**, 229–234, https://doi.org/10.5194/piahs-370-229-2015.

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

Willmott, C. J., and K. Matsuura, 2005: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Res.*, **30**, 79–82, https://doi.org/10.3354/cr030079.

Wu, T., W. Zhang, X. Jiao, W. Guo, and Y. A. Hamoud, 2021: Evaluation of stacking and blending ensemble learning methods for estimating daily reference evapotranspiration. *Comput. Electron. Agric.*, **184**, 106039, https://doi.org/10.1016/j.compag.2021.106039.

Zeng, Z., W. W. Hsieh, A. Shabbar, and W. R. Burrows, 2011: Seasonal prediction of winter extreme precipitation over Canada by support vector regression. *Hydrol. Earth Syst. Sci.*, **15**, 65–74, https://doi.org/10.5194/hess-15-65-2011.