IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 25 July 2023, accepted 14 August 2023, date of publication 28 August 2023, date of current version 13 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3309054

APPLIED RESEARCH

Machine Learning Approaches to Improve North American Precipitation Forecasts

CENKER SENGOZ[1], SHEELA RAMANNA[1], SCOTT KEHLER[2], RUSHIL GOOMER[1], AND PAUL PRIES[2]

[1]Department of Applied Computer Science, The University of Winnipeg, Winnipeg, MB R3B 2E9, Canada
[2]Weatherlogics Inc., Lorette West, MB R5K 0A6, Canada

Corresponding author: Sheela Ramanna (s.ramanna@uwinnipeg.ca)

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Alliance under Grant ALLRP-568786-21.

ABSTRACT Numerical weather prediction (NWP) is a challenging task which involves working with micro and macro-scale spatio-temporal parameters susceptible to biases and accuracy problems. In recent years, machine learning has grown in popularity with the increasing demand in accurate weather predictions. In this study, we adopt a multimodel (ensemble) forecasting approach by collecting precipitation data from multiple NWP models of Canadian, American and European weather agencies in an effort to deploy an optimal machine learning-based weather model for real-time precipitation forecasting that will outperform the baseline. We considered 8 NWP models as inputs and combined them to create ensemble predictors using 5 different machine learning techniques along with a baseline model (mean of eight input NWP models). We demonstrate that machine learning approaches can improve upon the results of the individual NWP models. The best results were obtained by the neural-network variants with 17% improvement in the mean absolute error, 3% in the root mean squared error, 47% in the median absolute error, 5% in the maximum error, 70% in the relative bias, 41% in the false alarm ratio and 8% in the critical score index over the baseline. Neural networks also complied with the practicality constraints, with minutes of training time and near-real time prediction time.

INDEX TERMS Machine learning, neural networks, regression, numerical weather prediction, precipitation forecasting.

I. INTRODUCTION

Weather monitoring and prediction have always been an integral part of ensuring the safety and preparedness of our daily lives. Numerical Weather Prediction (NWP) [1] refers to using mathematical models to process weather data to make forecasts. One challenging target for NWP is precipitation [2]. Precipitation is often parameterized using functions for cloud dynamics and micro-physics, requiring the model to account for extensive spatial and temporal scales.

To remove uncertainties and improve upon NWP models, researchers have developed several strategies, the most common being employing ensembles of multiple models [3], [4], [5], [6]. [7] studied the effect of NWP ensemble sizes on its prediction accuracy, having different physical parameters as inputs to the same NWP model. References [8] and [9] used bayesian post-processing techniques discussed in [10] to combine multiple NWP forecasts reducing errors in rainfall prediction. Reference [11] evaluated the use of an ensemble of seven independent NWP models and compared them to its individual members, over Australia. The ensemble models reduce the uncertainties of a single model but fail to map the non-linear relationships between the model output and real-world observations.

In recent years, machine learning and deep learning techniques have shown remarkable success in various domains and have the potential to improve NWP models. Artificial

The associate editor coordinating the review of this manuscript and approving it for publication was Szidonia Lefkovits.

neural networks can combine information from discrete sources to generate precipitation predictions [12], [13], [14]. One of the key advantages of using machine learning techniques is its ability to automatically extract features from vast amounts of data, without the need for manual feature engineering. Casper Kaae et al. [15] introduce the Metnet model that uses convolutional neural networks (CNNs) to combine data from GOES-16 satellite and ground radars to predict precipitation up to 8 hours. This can greatly improve the accuracy of weather predictions by incorporating more relevant information and capturing complex atmospheric processes that traditional NWP models might miss. References [16], [17], and [18] conducted surveys of various machine learning and deep learning algorithms being used to predict rainfall. These studies examined the use of different model architectures along with their profuse input data.

Reference [19] introduces a novel approach combining the U-net model with PredRNN architecture [20] to improve their time complexity and reduce errors for precipitation forecasting. Furthermore, [21] compares the ability of one-dimensional CNNs to predict monthly rainfall over Innisfail with climate indexes measured over oceans as input. In [22] authors introduce a GRU architecture with a self-attention mechanism focusing on high-impact weather events such as hurricanes using radar precipitation inputs. Ravuri et al. [23] present a GAN architecture with spatial and temporal discriminators, taking in 20 mins of ground radar data as inputs and predicting precipitation for the next 1.5 hours. The agile nature of these learning algorithms can handle non-linear relationships between variables and make predictions based on patterns in historical data, which can enhance the performance of NWP models'.

Post-processing individual NWP models or their ensembles using machine learning techniques [24], [25], [26], [27] could remove the model bias and uncertainties in the output. A considerable amount of research has been done in this field, however, most researchers fail to deploy these models effectively for higher precipitation levels. Fan et al. conducted a comparative study [28] to determine the best way to combine satellite imagery, rain gauges, and ECMWF reanalysis forecasts. A variety of algorithms were explored, including random forests (RF), long-term short memory (LSTM), fully connected neural networks (FCNN), and linear regression (LR). All the approaches reduce error in moderate precipitation conditions but struggle during heavy rainfall. Out of these, LSTM's were reported to be the most robust way of prediction. Frnda et al. [29] and Zhou et al. [30] created FCNN and U-net architectures respectively to post-process and improve European Centre for Medium-Range Weather Forecasts (ECMWF) model by extracting features from various environment variables and weather indexes. Reference [29] claims to have a 45% improvement in RMSE value (24 hours accumulated precipitation) and [30] reports an improved threat score for 0.1 mm, 3 mm, 10 mm, and 20 mm precipitation depths by 19.7%, 15.2%, 43.2%, and 87.1%, respectively (72 hours accumulated precipitation). Ko et al. in [31] created an Extreme Gradient Boosting (XGboost) model to improve predictions provided by the Korea Meteorological Administration. Their proposed algorithm provides 3-hourly accumulated precipitation forecasts by using environment variables such as precipitation values, wind speed, and humidity.

In [32] researchers used wavelet transformation along with machine learning methods, downscaling NWP forecasts to bias correct seasonal precipitation values. They report 21-33% reduced root mean square error (RMSE), indicating good performance in the bias correction. Research by Vladimir et al. [33] used an ensemble of nonlinear neural networks to improve 24-hourly precipitation forecasts over the Continental US. Eight different NWP models were used as inputs to ten independent neural networks (NN), the results of which were then averaged. Comparisons are made between these averaged results to the results obtained by human forecasters, and the NN multimodel ensemble was as accurate as human forecasts. The results show that the NN ensemble improves upon the pre-processed NWP models and reduces high bias at low precipitation and low bias at high precipitation levels. In [34] the google research team expanded the capabilities of their previously discussed weather forecasting model MetNet, by taking a hybrid approach and combining the state-of-the-art high-resolution rapid refresh (HRRR) NWP model with satellite and ground radar precipitation data using a ConvLSTM. The proposed approach can effectively predict precipitation up to 12 hours and outperform the HRRR predictions on various categorical metrics.

Weather forecasts for over 24 hrs present even more challenges, as discussed by Fan et al. in [35]. Their study focused on improving Week 3-4 precipitation and air temperature predictions using different neural network configurations. Although neural networks show promising improvements, they are still heavily dependent upon the initial NWP predictions.

The objectives of this study are as follows: i) to adopt a multimodel (ensemble) forecasting approach using precipitation data from multiple NWP models of Canadian, American and European weather agencies, ii) to identify and experiment with several light-weight machine learning algorithms to be used as merging methods, and iii) to train and deploy an optimal machine learning-based weather model for real-time precipitation forecasting that will outperform the baseline.

In order to avoid confusion in terminology between the NWP models that are used as inputs and the machine learning models that were trained, the former will be referred to as *weather models (WMs)*, whereas the latter will be referred to as *trained ML models* in this paper.

The aforementioned objectives were realized by collecting and extracting daily accumulated precipitation data covering most of the continental USA and Canada, performing extensive preprocessing and feature selection (overviewed in figure 1), training five well-known ML models and

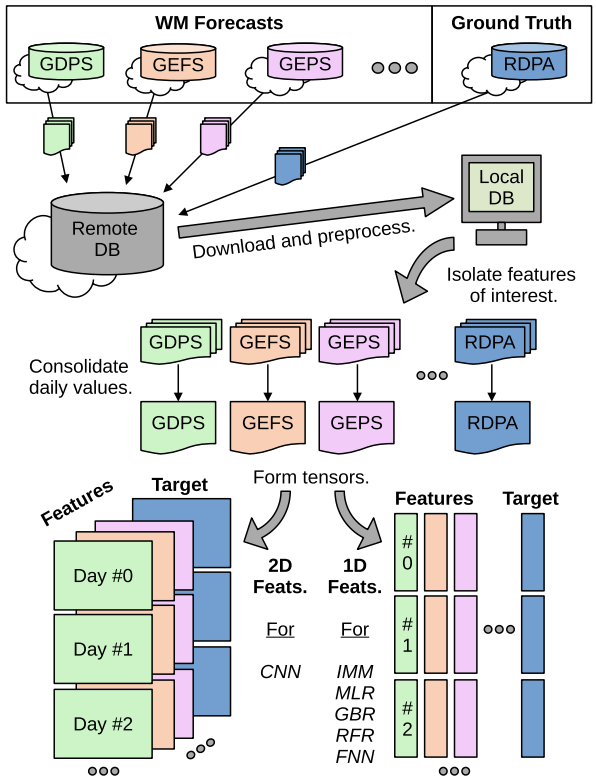


FIGURE 1. Data acquisition & preprocessing pipeline.

analysing results using nine well-known metrics. The NWP precipitation forecasts (input features) were derived from the following WMs: The Global Deterministic Prediction System (GDPS), Global Ensemble Forecast System (GEFS), Global Ensemble Prediction System (GEPS), Global Forecast System (GFS), Icosahedral Nonhydrostatic (ICON), North American Mesoscale Forecast System (NAM), Regional Deterministic Prediction System (RDPS), and the Regional Ensemble Prediction System (REPS). An Input Means Model (IMM) which is the simple arithmetic mean of all eight WMs was used as a baseline. In addition, several secondary features of precipitation data were also examined.

The practicality of a proposed approach would also have to be taken into account for our use case. Currently, the baseline technique is commissioned and its processing overhead is trivial. The improved approach would similarly need to be resource-light with manageable memory and processing time. In particular, we aimed for solutions which can be deployed on a desktop computer with tensorflow-capable GPU, with a memory consumption of under 16 GB and a processing time of under one hour for training and near real-time for prediction.

Our study revealed that while all ML models introduce improvements, the neural network models showed superior performance across most metrics. With fully-connected neural networks, we obtained 17% improvement in the mean absolute error, 3% in the root mean squared error, 47% in the median absolute error, 5% in the maximum error, 70% in the relative bias, 41% in the false alarm ratio and

FIGURE 2. The region of study.

8% in the critical score index over the baseline, while maintaining <2% difference in correlation coefficients and probability of detection. Neural networks also complied with the practicality constraints, with minutes of training time and near-real time prediction time.

The unique contributions of this work are i) a pairwise correlation analysis between the estimated precipitation and a range of weather and spatio-temporal features from eight WMs, and ii) a comparative analysis of the WMs with their precipitation forecasts against five ML models and a baseline in terms of nine identified metrics.

The rest of the paper is organized as follows: In section II, we give details of the geographical area, input weather models' details, secondary features and feature selection process covered in this study. In section III, we discuss data preprocessing, implemented ML models and the metrics used to assess the performance the ML models. In section IV, we present detailed analysis of our experiments in terms of each of the nine metrics. In section V, we present the overall performance of the five trained ML models as well as the input WMs in terms of the nine evaluation metrics used in this study, followed by concluding remarks in section VI.

II. DATA

A. AREA OF STUDY

We consider a geographical area that covers the majority of the continental USA and Canada, along with the surrounding region, as shown in figure 2. Specifically, it is confined to the 24th parallel from the south, 70th parallel from the north, 218th meridian from the west, and 308th meridian from the east. The area is gridded at an increment of 0.25 degrees, starting from the origin at the 24th parallel and 218th meridian. This results in 369 rows and 721 columns, 266,049 grid units for which to forecast and gauge the daily accumulated precipitation.

B. INPUT WEATHER MODEL DETAILS

Table 1 presents the spatial and temporal attributes of the samples of each of the WMs as well as their original

TABLE 1. Input weather models' properties.

WM	Spatial Resolution	Temporal Period	Original Accumulation Period
GDPS	0.25°	24 h	Running total
GEFS	0.25°	24 h	6-hourly
GEPS	0.25°	24 h	6-hourly
GFS	0.25°	24 h	Running total
ICON	0.25°	24 h	Running total
NAM	0.25°	24 h	12-hourly
RDPS	0.25°	24 h	Running total
REPS	0.25°	24 h	Running total

accumulation periods for precipitation forecasts. The grid spacing in table 1 represents the WMs after they were regridded for this study. The WMs natively use various different grid spacings, so they were regridded to a common grid to make processing easier. The feature of interest is the daily accumulated precipitation which is considered as a primary input feature as well as the target output for the ML models. Dozens of other variables ranging from visibility to soil temperature and convective available potential energy are also predicted by different subsets of these WMs. However, many of them are intermittent across the time and space domains so they are not available for wide-scale experimentation. We will, however, consider the persistent ones as potential secondary features. In the following subsections, we present a brief description of each of the WMs considered.

1) GDPS

The Global Deterministic Prediction System (GDPS) [36] is a WM that is used for global data assimilation and medium range forecasting. It is developed by the Meteorological Service of Canada (MSC) at the Canadian Meteorological Centre (CMC). The version that is used in this study (v8.0) was released in December, 2021. It provides forecasts two times a day for a lead time of ten days with three-hourly increments. The forecasts are made on Yin-Yang horizontal grid with a horizontal grid spacing of 0.135 degrees (15 km). It covers a range of variables including precipitation, wind gusts, humidity, cloud cover, temperature, wind speed and wind directions.

2) GEFS

The Global Ensemble Forecast System (GEFS) [37] is a WM created by the United States National Centers for Environmental Prediction (NCEP), a branch of National Oceanic and Atmospheric Administration (NOAA). It has a horizontal grid spacing of 0.25 degrees (25 km) and a forecast lead time of sixteen days (384 hours) with an output timestep of three hours. The forecasts are made four times a day. In our work, version 12.0 is employed, which was released in September 2020. Its suite of variables include temperature, humidity, wind speed and direction, precipitation and cloud cover amongst others. Unlike GEFS, this WM does not make a single deterministic forecast but rather, probabilistic forecasts based on a range of ensemble members each of which works with a marginally perturbed set of inputs, resulting in a probabilistic distribution to account for the intrinsic uncertainty of the weather conditions. This particular WM uses $30 + 1$ ensemble members (one is used for control) and we considered their mean output as the feature, in our study.

3) GEPS

The Global Ensemble Prediction System (GEPS) [38] is another WM developed by the MSC at CMC, Canada. Like GEFS, it is an ensemble WM. It has $20 + 1$ perturbed members. Its forecasts have a lead time of sixteen days, and the forecasts are executed two times a day with a timestep of three hours. It has a horizontal grid spacing of 0.35 degrees (39 km). The variables covered by this WM include precipitation, wind speed and direction, temperature and humidity. We used version (v7.0) which was released in December, 2021.

4) GFS

The Global Forecast System (GFS) [39] is a global WM created by the NCEP of the United States as part of its suite of numerical tools. It is widely used in the meteorological community and provides detailed forecasts of global weather conditions. It produces forecasts four times a day with a lead time of sixteen days and it has a horizontal grid spacing of 13 km. The first five days have one-hourly forecast periods and afterwards, it increases to three hours. We used version 16 which was implemented in March 2021. Its variables include wind gust, temperature, humidity, wind speed and direction, precipitation and cloud cover.

5) ICON

ICON (short for the Icosahedral Nonhydrostatic model) [40] is a WM developed by the German weather service, Deutscher Wetterdienst (DWD). This is a global model that uses an icosahedral grid, which is a type of grid that is based on a geometric shape with 20 faces, to represent the earth's surface. The actual global grid is finer, comprised of 2,949,120 triangles and it amounts to a mesh size of 13 km. Forecasts are made four times a day, with a lead time of 180 hours for the runs at 00 and 12 UTC. For the 06 and 18 UTC runs, the lead time is 120 hours. For the first 78 hours, output period is one hour after which it increases to three hours. The variable that was available from this WM was total precipitation forecast.

6) NAM

The North American Mesoscale Forecast System (NAM) [41] is a WM that provides forecasts for the United States, Canada, and Mexico. It is developed by the NCEP of the United States and it uses a high-resolution model to provide detailed forecasts. The NAM model is typically used for short-range weather forecasting and to support decision-making in

industries such as aviation, energy, and transportation. This WM runs 4 times a day and makes forecasts with a lead time of 84 hours with a forecast timestep of 3 hours. Its horizontal grid spacing is 12 km. The suite of variables covered by this WM include wind gust, temperature, humidity, wind speed and direction, and precipitation. The timestep is 3 hours.

7) RDPS

The Regional Deterministic Prediction System (RDPS) [42] is developed by the Canadian Meteorological Centre to produce detailed weather forecasts for Canada and the United States. It operates on a Limited Area Model (LAM) grid with a size of 1108 by 1082 and a horizontal grid spacing of 0.09 degrees (10 km). We used version 8, which was released in December 2021. Predictions are made 4 times a day and the forecast lead time is 84 hours. The timestep for the forecasts is 300 seconds. The set of variables it supports include precipitation, wind gust, humidity, cloud cover, temperature, wind speed and direction.

8) REPS

The Regional Ensemble Prediction System (REPS) [43] is the ensemble counterpart of RDPS. It is likewise developed by the CMC and it uses the same grid as RDPS, with a horizontal grid spacing of 0.09 degrees (10 km) covering Canada and United States. Version 4 is used, which was released in December 2021. The ensemble consists of 20 + 1 members. It runs 4 times a day with a forecast lead time of 72 hours. The timestep is 300 seconds. Amongst the forecast variables delivered by the model are precipitation, humidity, temperature, wind speed and direction.

C. RDPA - GROUND TRUTH TARGET

We used the Canadian Regional Deterministic Precipitation Analysis System (CaPA-RDPA) [44] from CMC, as precipitation estimates to represent the ground truth for our precipitation forecasts. This system works on the same grid for the RDPS, covering the United States and Canada. The analyses are executed four times a day (00, 06, 12, 18Z), producing estimates for the preceding six-hour window. The grid spacing is 10 km. RDPA version 5.2.0 is used.

D. FEATURE SELECTION

Alongside daily accumulated precipitation (PR), we also considered twelve secondary features that are present in our data set. The features are summarized in table 2. These include WM-bound features like 2 m air temperature (TM), 2 m relative humidity (RH), U-component of the wind at 10 m above ground (UW), V-component of the wind at 10 m above ground (VW), total cloud cover percentage (CL), wind gusts at 10 m above ground (GS); the elevation of the ground surface (EL); the spatial features of latitude (LT) and longitude (LN); and the temporal features in forms of varying representations of julian date (JD).

TABLE 2. List of considered input features. Cardinality shows the number of input WMs that can produce this feature. For WM-agnostic geographical and spatio-temporal features it defaults to (1).

Feature	Shorthand	Unit	Cardinality
Daily precipitation	PR	kg / m^2	8
2 metre air temperature	TM	K	7
2 metre relative humidity	RH	%	7
10 metre U wind component	UW	m / s	7
10 metre V wind component	VW	m / s	7
10 metre wind gust	GS	m / s	4
Total cloud coverage	CL	%	4
Elevation	EL	m	(1)
Latitude	LT	°	(1)
Longitude	LN	°	(1)
Julian date	JD	-	(1)
Julian date (cosine)	Cos(JD)	-	(1)
Julian date (sine)	Sin(JD)	-	(1)

TABLE 3. List of considered feature aggregations.

Aggregation Method	Features to consider (per WM)	# Features
Daily mean	PR, TM, RH, UW, VW, GS, CL	44
Daily std	PR, TM, RH, UW, VW, GS, CL	44
Mean of daily means	PR, TM, RH, UW, VW, GS, CL	7
Mean of daily std	PR, TM, RH, UW, VW, GS, CL	7
Std of daily means	PR, TM, RH, UW, VW, GS, CL	7
Std of daily std	PR, TM, RH, UW, VW, GS, CL	7
Unary (spatial)	LT, LN,	2
Unary (temporal)	JD, cos(JD), sin(JD)	3
Unary (geographical)	EL	1
Gauged Precip.	Ground truth	1
Total		123

Table 2 also shows the cardinality of each feature. The primary feature of precipitation is available in every WM whereas the secondary ones are only available in a subset of them. In any event, because the number of WM-bound features to be considered are proportional to the number of WMs that produce them, experimenting with all of their combinations would prohibitively increase the complexity of the input space. Furthermore, whether they carry the potential to improve the ML models would still have to be assessed, particularly for the secondary features. We therefore opted for a correlation analysis to preview their prospects.

Specifically, we explored pairwise correlation of RDPA (ground truth) against 123 potential feature forms, summarized in table 3. These feature forms include the daily mean (average of eight 3-hourly values) and standard deviation of individual WM forecasts as well as the aggregated values of multiple WMs (e.g. cross-model means of daily mean forecasts). Note that daily mean and daily accumulated precipitation are fully correlated so they can be used interchangeably for the purpose of this analysis.

The results of the correlation analysis over a sample set of 3 months is shown in table 4. As observed, the most correlated features against the RDPA are the different forms of daily precipitation features, dominating the top of the table, with all of the 8 WMs having a correlation value of more than 0.85 for daily mean.

All WM-bound secondary features lag significantly behind with the best performing one, mean cloud coverage by RDPS, situated well below 0.5. The geographical and spatio-temporal ones are yet to show any correlation.

TABLE 4. Absolute correlations of 123 feature aggregations against the RDPA ground truth target. Temporal window: (Jan 2022 - Apr 2022).

Feature	Corr.	Feature	Corr.	Feature	Corr.	Feature	Corr.
<Ground Truth - RDPA>	1.00	GS std (NAM)	0.34	RH std of mean	0.20	TM std (RDPS)	0.07
PR mean (REPS)	0.94	UW std (GEPS)	0.34	VW mean (GDPS)	0.19	TM std (GEPS)	0.07
PR mean of means	0.93	VW std (NAM)	0.34	VW mean (NAM)	0.19	UW mean (NAM)	0.07
PR mean (RDPS)	0.93	VW mean of std	0.33	VW mean of means	0.19	LT	0.06
PR std (REPS)	0.91	GS mean of means	0.33	VW mean (REPS)	0.19	sin (JD)	0.05
PR mean (GDPS)	0.91	VW std (GEFS)	0.33	VW mean (GEFS)	0.19	UW mean (GFS)	0.04
PR mean (GEPS)	0.91	UW std (REPS)	0.33	VW mean (RDPS)	0.19	UW mean of means	0.04
PR std (RDPS)	0.91	GS mean of std	0.32	VW mean (GFS)	0.19	UW mean (GEFS)	0.04
PR mean of std	0.90	VW std (RDPS)	0.32	CL std (GEFS)	0.19	UW mean (GEPS)	0.04
PR std (GDPS)	0.89	VW std (GDPS)	0.32	TM mean (REPS)	0.18	TM std (GEFS)	0.03
PR mean (GEFS)	0.88	VW std of std	0.32	TM mean (RDPS)	0.18	UW mean (REPS)	0.03
PR mean (ICON)	0.87	GS mean (GFS)	0.31	VW mean (GEPS)	0.18	UW mean (GDPS)	0.03
PR mean (GFS)	0.86	RH mean (GDPS)	0.31	TM mean (GDPS)	0.18	UW mean (RDPS)	0.03
PR std (ICON)	0.85	VW std (GEPS)	0.30	TM mean (GEPS)	0.18	JD	0.03
PR mean (NAM)	0.85	GS mean (RDPS)	0.30	TM mean of means	0.18	cos (JD)	0.02
PR std of std	0.82	GS mean (GDPS)	0.30	TM mean (GEFS)	0.17	CL std (RDPS)	0.02
PR std of mean	0.80	RH mean (RDPS)	0.30	VW std of means	0.16	RH std (RDPS)	0.02
PR std (GEPS)	0.77	VW std (REPS)	0.30	TM mean (NAM)	0.16	RH std (NAM)	0.02
PR std (GEFS)	0.75	GS std (RDPS)	0.29	UW std of means	0.16	RH std (GDPS)	0.02
PR std (GFS)	0.73	RH mean (REPS)	0.29	TM mean (GFS)	0.15	TM std of means	0.02
PR std (NAM)	0.68	GS std (GDPS)	0.29	CL mean of std	0.13	RH std (GEFS)	0.02
CL mean (RDPS)	0.45	UW std of std	0.28	TM std of std	0.13	RH std (GFS)	0.02
CL mean (GDPS)	0.44	CL mean (GEFS)	0.28	RH mean (GEFS)	0.13	RH mean of std	0.01
CL mean of means	0.38	GS std (GFS)	0.28	RH mean (GFS)	0.12	RH std (REPS)	0.00
GS mean (NAM)	0.38	CL mean (GFS)	0.26	CL std of std	0.10	RH std of std	0.00
UW std (NAM)	0.37	RH mean (GEPS)	0.26	EL	0.09	RH std (GEPS)	0.00
UW std (GFS)	0.36	RH mean of means	0.25	LN	0.09	CL std (GDPS)	0.00
UW mean of std	0.36	CL std of mean	0.22	TM std (NAM)	0.09		
UW std (GEFS)	0.36	RH mean (NAM)	0.22	TM std (GDPS)	0.08		
UW std (GDPS)	0.35	GS std of mean	0.21	TM std (REPS)	0.08		
UW std (RDPS)	0.35	CL std (GFS)	0.20	TM mean of std	0.07		
VW std (GFS)	0.34	GS std of std	0.20	TM std (GFS)	0.07		

Additionally, preliminary training of the ML models suggested that the secondary features did not result in any significant contribution that would justify their inclusion, given that they increase the complexity of the input space. Consequently, we only employed the daily accumulated precipitation values as features against our target value of daily precipitation.

III. METHODS

In this section, we discuss the data preprocessing steps taken and the machine learning models that we used for precipitation forecasting, along with the baseline regressor.

A. DATA PREPROCESSING

As the WMs of interest and the precipitation estimates (RDPA) come from different weather service providers, their data has to be spatially and temporally aligned before it can be used. This preprocessing phase involves extracting the field of interest (i.e. daily accumulated precipitation), such that the type of map projection, covered area, dates and times are all consistent. We acquired the data in GRIB2 format and regridded them to cover our region of interest using wgrib2. The WM forecasts were in 3-hourly formats, and the RDPA data was in 6-hourly format. Both types of data were preprocessed with the help of numpy, pandas, and pygrib for Python[1] and aggregated into 24-hourly blocks. It should be noted that our daily cycle is based on Central Standard Time Zone, which covers from UTC+06 to UTC+30.

Our data acquisition and preprocessing pipeline is summarized in figure 1. To enable our dataset for use by our machine learning models, we converted it into a tensor form. For every ML model, we created a 2D tensor where each column represented a forecast by an individual WM. This was done by flattening and concatenating daily grid forecasts of respective WMs. In this setup, each row constitutes a data point and the rows can be shuffled without regards to latitude, longitude and date. The last column is used for the RDPA ground truth data as the target.

The only model which uses a different layout is the convolutional neural network architecture. It utilizes a 3D tensor where each slice is the 2D grid forecast by a given WM, concatenated on a daily basis. In this case, a data sample is composed of the entire daily 2D grid forecasts by our input WMs rather than a single grid cell.

Throughout the experiments, the data set was further cleaned and enhanced as necessary by preprocessing methods such as removing the rows with missing cells, removing the days with missing data or input spectrum normalization, depending on the merging method.

B. MERGING METHODS

This section gives an overview of the five machine learning algorithms which were used to train the ML models.

[1] https://www.python.org/

1) BASELINE - INPUT MEANS MODEL (IMM)

The Input Means Model (IMM) is the baseline approach we used to combine the forecasts of multiple WMs in order to produce a single, more accurate prediction. This averaging technique for predictions is often used in data analysis, when there are multiple inputs available for making a prediction about a given data point. To implement this, we first take a set of 8 different predictors for the same data point and calculate the mean of these predictions as the final prediction as shown in Eqn. 1

$$P_{\text{IMM}} = \frac{\sum_{i=1}^{N} p_i}{N} \tag{1}$$

where $N = 8$ is the number of WMs to be input, p_i is the prediction of the i^{th} WM and P_{IMM} is the final prediction. This technique can be a useful method for improving the accuracy of predictions, particularly in case the different WMs being combined have different strengths and weaknesses. However, it is important to recognize that the technique relies on the assumption that the predictions of the different WMs are unbiased and that they are all equally valid. If this assumption is not met, the technique may not produce accurate results.

2) MULTIPLE LINEAR REGRESSION (MLR)

Because numerical weather prediction is a type of regression task, the next approach we considered is multiple linear regression (MLR). It is a widely used statistical technique used for making predictions about a continuous dependent variable denoted by (P_{MLR}) based on multiple independent variables denoted by (p_i) as shown in Eqn.2 assuming a linear relationship between the dependent and independent variables.

$$P_{\text{MLR}} = \beta_0 + \sum_{i=1}^{N} \beta_i \cdot p_i \tag{2}$$

In our case, $N = 8$ WM predictions for precipitation were used as inputs to the MLR algorithm to fit a linear regression model, involving the estimation of coefficients (β_i) and the intercept (β_0). MLR is particularly useful for bias correction and is also considered resource-light.

We used non-normalized input as normalization produced suboptimal results for this particular model and the data set.

Regression is performed using ordinary least squares method and it is the fastest and least complex technique that we explored apart from the baseline (input means) model. In our experiments, it consistently took well under a minute to fit MLR to our data.

It should, however, be noted that this model still assumes linear associations between the input and the output data. It also works best if the residuals (errors) are normally distributed. Thus, MLR may or may not be the a good option given these assumptions.

3) GRADIENT BOOSTING REGRESSION (GBR)

Gradient boosting regression (GBR) is an ensemble method that incorporates the predictions of multiple weak models, such as decision trees, in order to produce a superior, unified predictor. Gradient boosting of regression trees produce competitive, highly robust, interpretable procedures for both regression and classification, and is especially appropriate for mining less than clean data [45]. Training is performed in a sequential, iterative fashion where each tree is constructed to reduce the errors of its predecessors. Unlike MLR, a GBR model is non-linear by nature so it can learn and represent more complex relationships between the input variables and the output, and it can produce more accurate results than individual WMs.

The potential drawback of this model is that it can be computationally intensive and is also sensitive to the hyper parameters used to fit the model, so careful tuning is necessary to achieve optimal performance. Accordingly, we employed the histogram-based version of the gradient boosting regression method, which reduces the computational complexity and memory requirements of a traditional GBR approach by orders of magnitude through placing the data points into so-called ''bins'', discretizing and the input space.

The model was trained via a squared error loss function and a learning rate of 0.1 creating up to 31 maximum leaf nodes and a maximum tree depth of 7 for each regression tree. The input space is reduced to 256 bins. We employed 100 trees (amounting to 100 iterations). The convergence of the model took about two minutes. Tree-based algorithms typically don't require input normalization, hence the inputs are used in their native spectrum.

4) RANDOM FOREST REGRESSION (RFR)

Random forest regression (RFR) is the other machine learning technique [46] involving decision trees that we considered. It is an ensemble method that combines the predictions of the trees, each of which is trained on an arbitrarily selected subset of the training samples.

Like GBR, RFR is developed to capture non-linear complex relationships between the input features and the output variable. It is able to handle large datasets with many features and it is relatively resistant to overfitting, which makes it a robust choice for many prediction tasks. Unlike GBR, each tree in the ensemble is trained independently so they can be concurrently trained.

RFR may require training a large number of decision trees and may be time consuming. It is likewise sensitive to the hyperparameters used to fit the model.

In our experiments, we utilized a set of 10 trees, each representing a separate estimator, with a maximum depth set to 7. The loss function os based on squared error, like the preceding methods. The execution time for the concurrent operation was approximately 5.2 minutes. Using a handful of trees enabled the entire training process to be completed concurrently on a modern desktop computer. In addition, preliminary experiments showed that using up to 100 trees increased the execution time without a noticeable change

in the performance. Therefore, this approach emerged to be more advantageous. Input is not normalized, with the model being tree-based.

5) FEEDFORWARD AND CONVOLUTIONAL NEURAL NETWORKS

As far as the machine learning techniques go, neural networks opened a new realm of possibilities. Influenced by the biological brains, they form models consisted of interconnected nodes or ''neurons'', each of which can ''fire'' in accordance to their activation functions, transforming the input as propogating them through the network into desired outputs. They have found a wide range of application fields and weather forecasting may as well be one of them.

In this work we considered them in two forms:

- Feedforward neural networks (FNN), which can be regarded as vanilla form, employ layered and unidirectional network connections. This version can handle each data point as individual inputs. It doesn't maintain a structural information which gives us flexibility on shuffling the data and disregard the spatio-temporal information altogether. On the flipside, one might expect a degree of correlation amongst neighbouring regions which may suggest such disregarded information may cause the model to be suboptimal.
- Convolutional neural networks (CNN) on the other hand, are designed to handle two-dimensional data with image-like layouts. This type of network also has layered structure but instead of transforming the input data points individually, it uses predefined set of kernels as filters to convolve the input images. The obvious advantage is that the model is able to retain and leverage spatial information. On the other hand, the number of samples available for training is orders of magnitude less, what used to be 266,049 separate data points is now represented as a single sample, a 2D grid of shape 369×721.

We considered a range of hyperparameters for both neural net based models in our preliminary experiments. They include various number of hidden layers {1, 2, 3, 5, 10}, architectures {uniform-width, widening, narrowing, hourglass}, initializations {random normal, random uniform, zeros, ones}, optimizers {SGD, Adam, RMSProp}, loss functions {squared, absolute, squared logarithmic} number of neurons in hidden layer {1, 5, 10, 100, 1000}, dropout regularization rates {0.1, 0.2, 0.5}, kernel filter counts {8, 16, 32, 64}, and shapes {(3×3), (5×5), (7×7)}. The configurations yielding the optimal results that we observed are summarized in table 5. Visualizations of those neural network architectures are also provided, in figures 3 and 4, respectively. Training time for both types of networks are around 1 minute until convergence, excluding the overhead for preprocessing. Neural networks expect normalized input, therefore the data is standardized via z-score normalization before being fed to either neural network.

TABLE 5. Trained ML model hyperparameters yielding the most optimal results that we observed.

Approach	Specifications
IMM	Loss: N/A Normalized: N/A
MLR	Loss: squared error Normalized: No
GBR	Number of trees:100 Tree depth: 7 Loss: squared error Normalized: No
RFR	Tree depth: 7 Number of trees: 10 Loss: squared error Normalized: No
FNN	Number of trainable params: 109,201 Number of hidden layers : 2 Number of neurons in hidden layer : (1000, 100) Dropout: 0.1 Optimizer: RMSprop Kernel initialization: Random normal Batch size: 200,000 Loss: squared logarithmic error Normalized: Yes (Z-Score)
CNN	Number of trainable params: 79,105 Number of hidden layers: 3 Dropout: 0.1 Optimizer: RMSprop Kernel initialization: Random normal Kernel count and shape: $64 \times (3 \times 3)$ Batch size: 1 Loss: squared logarithmic error Normalized: Yes (Z-Score)

C. METRICS

In this section, we introduce the performance metrics that we used.

In this project, we considered a number of metrics to evaluate our trained ML models. They are summarized in table 6. They can be elaborated as follows:

- Mean Absolute Error (MAE) [47] provides an understanding on the average error each model makes across all observations. It is defined as the average absolute difference between the predicted and actual values, where n is the number of observations, $\hat{y}_i$ is the predicted output and y_i is the actual output for the i^{th} observation.
- Root Mean Squared Error (RMSE) [47] penalizes larger errors. It is defined as the square root of the mean squared error (MSE), where n is the number of observations, $\hat{y}_i$ is the predicted output and y_i is the actual output for the i^{th} observation. RMSE is the average squared difference between the predicted and actual values.
- Maximum Error (MaxE) provides an insight into the worst-case performance scenario. It is defined as the maximum absolute difference between the predicted and actual values, where n is the number of observations, $\hat{y}_i$ is the predicted output and y_i is the actual output for the i^{th} observation.
- Median Absolute Error (MdAE) is a more robust measure with respect to outlier values. It is calculated

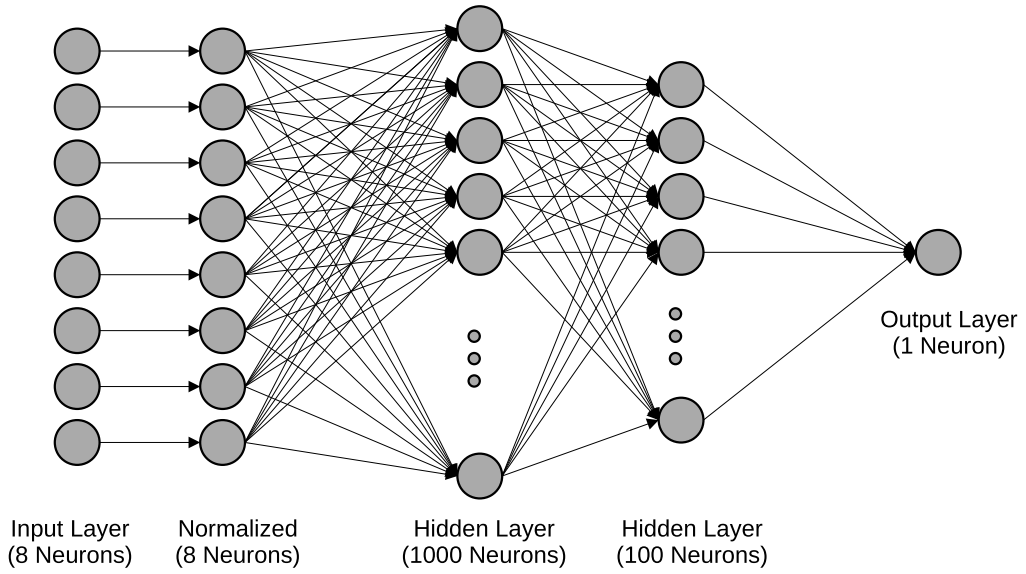


FIGURE 3. Feedforward neural network architecture.

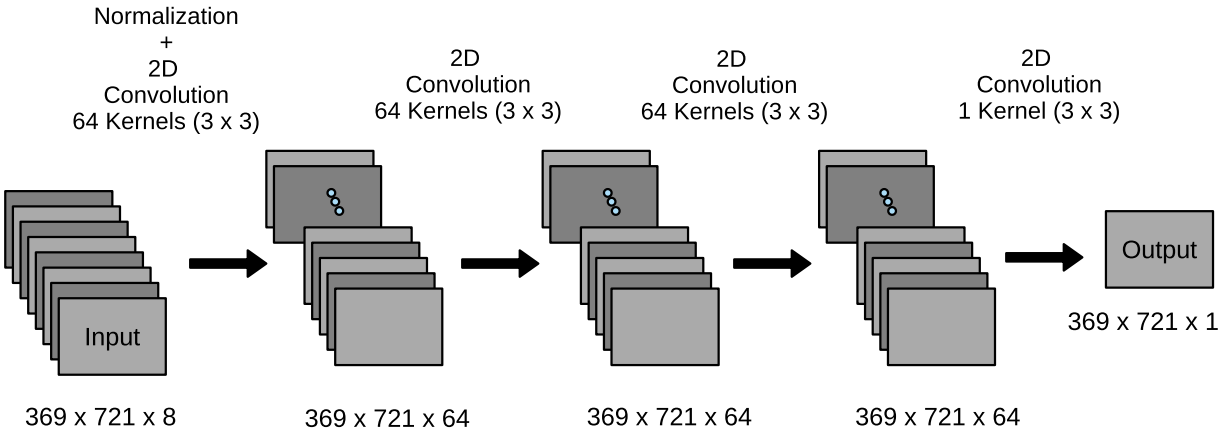


FIGURE 4. Convolutional neural network architecture.

by taking the median of the absolute differences between the predicted and actual values.

- Pearson Correlation Coefficient (CC) describes the level of association between the predicted and actual output values, where X and Y are predicted and actual output values respectively, Cov(X, Y) is the covariance of the two variables X, Y, Var(X) is the standard deviation of X and Var(Y) is the standard deviation of Y. A CC value of 1 indicates a strong positive relationship, a CC value of -1 indicates a strong negative relationship, and a CC value of 0 indicates no relationship at all.
- Relative Bias (RB) gives insight on whether a model tends to either over-estimate or under-estimate the output (e.g., precipitation value). It is defined as the average difference between the predicted and actual values relative to the mean of the actual values, where n is the number of observations, $\hat{y}_i$ is the predicted output and y_i is the actual output for the i^{th} observation.
- Probability of Detection (POD) measures how well the event of interest (precipitation) is detected. It is a measure of the ability of a classification model to correctly predict the presence of a particular class (in our case rainfall amounts). ''True Positives'' are the number of observations where both the predicted and actual values are same, and ''False Negatives'' are observations where a prediction wrongly indicates that an event did not occur, when in fact it did.
- False Alarm Ratio (FAR) measures the rate of erroneous precipitation forecasts. True Positives are the number of samples where both the predicted and actual values are same and False positives are instances where a test or prediction wrongly indicates that an event or condition has occurred.
- Critical Success Index (CSI) is a metric which essentially combines POD and FAR metrics into a single score. Specifically, it measures the ratio for the correctly predicted precipitation events to the sum of hits, misses and false alarms.

TABLE 6. Metrics.

Name	Formula	Unit
MAE	$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\lvert\hat{y}_i - y_i\rvert$	kg / m^2
RMSE	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$	kg / m^2
MdAE	$\text{MdAE} = \text{median}_{i=1}^{n}\lvert\hat{y}_i - y_i\rvert$	kg / m^2
MaxE	$\text{MaxE} = \max_{i=1}^{n}\lvert\hat{y}_i - y_i\rvert$	kg / m^2
CC	$\text{CC}(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$	/
RB	$\text{Relative Bias} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{\sum_{i=1}^{n}(y_i)}$	/
POD	$\text{POD} = \frac{\text{True Positives}}{\text{True Positives}+\text{False Negatives}}$	/
FAR	$\text{FAR} = \frac{\text{False Positives}}{\text{True Positives}+\text{False Positives}}$	/
CSI	$\text{CSI} = \frac{\text{True Positives}}{\text{True Positives}+\text{False Negatives}+\text{False Positives}}$	/

IV. SPATIAL AND TEMPORAL RESULTS

Our experimental data set spans the period from Dec 1, 2021 to June 7, 2022. In particular, we used roughly the first 4 months up to April 6, 2022 for training and the latter part for validation. We later analyzed the results in terms of all the 9 metrics introduced.

In order to get a better understanding of the outcomes, we studied the results in two separate contexts, spatial and temporal. For the spatial context, we considered our data grid of 369 by 721 cells. For each cell, we combined the results of the validation dates by taking their daily mean, resulting in a 2D map. Then we further consolidated every 14 cells to create a visually identifiable grid. For the temporal context, we combined the results across every cell in every day and created one daily data point along the time axis.

A. MAE - MEAN ABSOLUTE ERROR METRIC

Amongst the input WMs, the *lowest* value for the mean absolute error was achieved by REPS with 0.96 mm/day. It is followed by RDPS with a mean of 1.04. The poorest performance for this metric was observed for NAM at 1.31. The baseline input means model (IMM) was virtually the same with REPS at 0.95. For this MAE metric, we see that all the trained ML models improved upon the baseline IMM model, and the best models were the FNN and CNN neural networks with MAE of 0.79 and 0.78, respectively. For this metric, all the input WMs, and the trained ML models have similar tendencies in that, the geographical regions where they performed better or worse are mostly aligned.

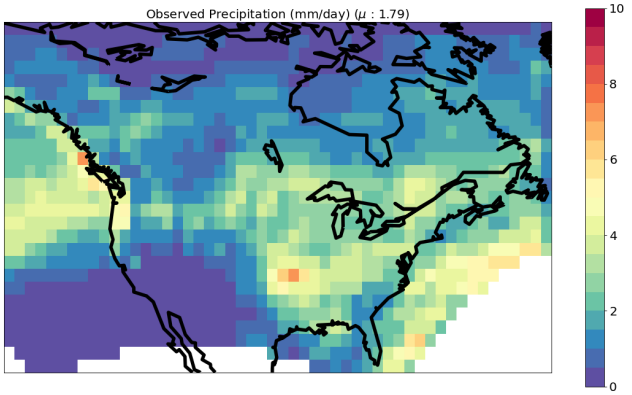


FIGURE 5. Observed (spatial) precipitation for the validation period.

We observe the *lowest* errors in drier zones of south-western US and the Pacific ocean. Higher errors were noticed in the eastern half of the US, and the Pacific coasts of the northwestern US and British Columbia. The improvements introduced by the machine learning algorithms in those areas are easily identifiable as smaller and lighter shades of purple, shown in figure 6. It is also notable that lower error values are observed at the drier parts of the spectrum (please see figure 5 for spatial precipitation distribution).

Figure 7 shows the mean absolute error over a temporal axis. We observe a similar pattern emerged where, on most days, the trained ML models outperformed the input WMs. Once again, NAM emerged as the poorest model, followed by GFS and ICON. Amongst the *best* performing trained ML models are FNN and CNN.

B. RMSE - ROOT MEAN SQUARED ERROR METRIC

In general, RMSE results are quite similar to MAE. The *best* performing input WM was REPS at 2.02 mm/day with the baseline IMM model at 1.95 mm/day shown in figure 8. The best overall result in terms of RMSE was from FNN and CNN trained ML models. Nonetheless, the performance order is not particularly the same as MAE. In fact, we see that for this metric, RDPS performed relatively poorer at 2.46, well behind GEPS at 2.23. Because RMSE is more sensitive to larger errors than MAE, this suggests that the input WMs have different tolerance for different kinds of errors. In terms of the trained ML models as well as the baseline IMM, we see noticeable improvements in all the models: MLR, GBR, RFR, FNN and CNN. Once again, the best performing are the neural network-based models (CNN and FNN) at 1.82, which is significantly lower than the input WMs. Temporally, we see the machine learning approaches lowering the RMS error as before, with the neural network models giving the lowest errors shown in figure 9.

C. MdAE - MEDIAN ABSOLUTE ERROR METRIC

Spatially, this metric yields a pattern much different than the previous ones as shown in figure 10. It is best optimized by FNN and CNN by a comfortable margin. Here, the baseline

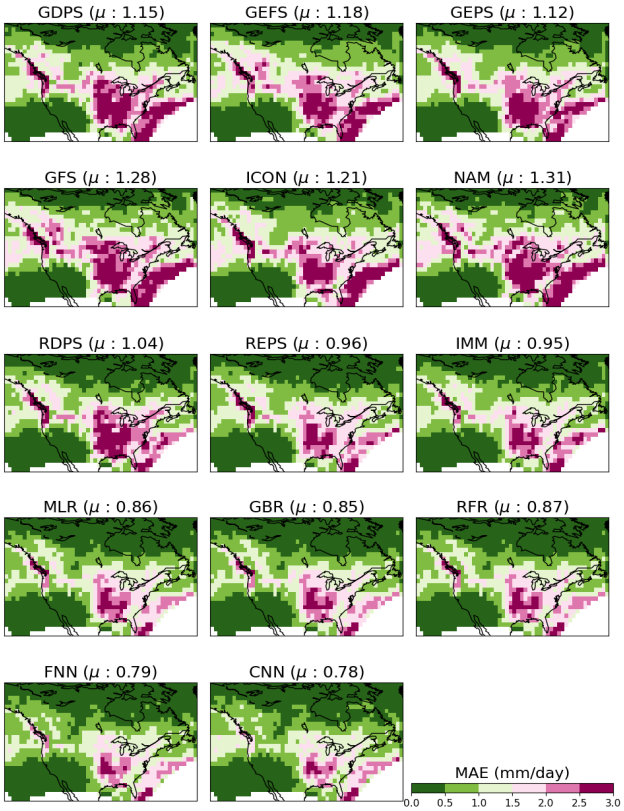


FIGURE 6. Mean absolute error (spatial). Lower is better.

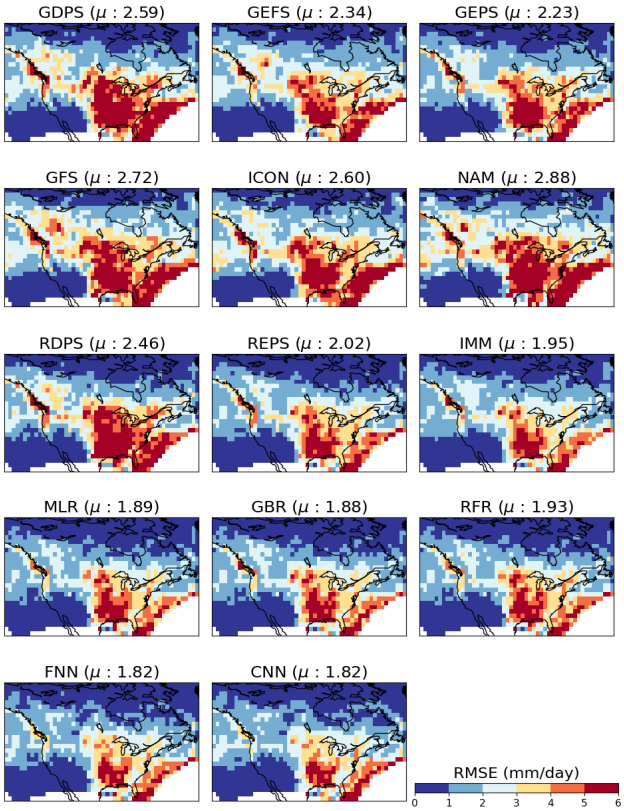


FIGURE 8. Root mean squared error (spatial). Lower is better.

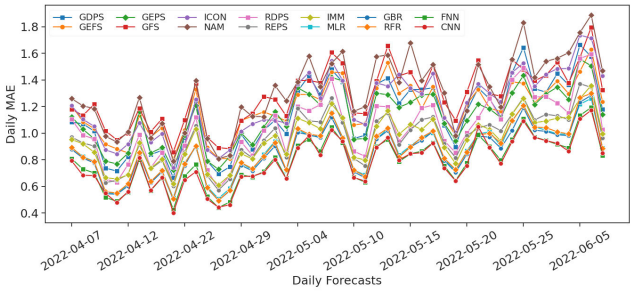


FIGURE 7. Mean absolute error (temporal). Lower is better.

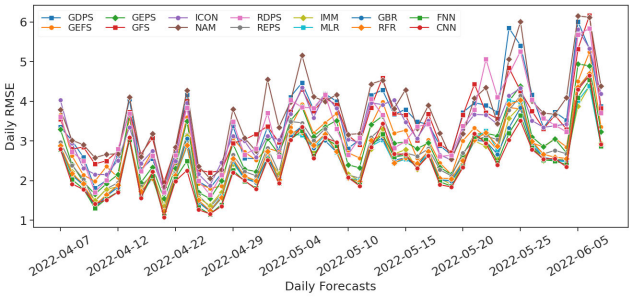


FIGURE 9. Root mean squared error (temporal). Lower is better.

IMM model fails to deliver any improvement, whereas the remaining 5 trained ML models perform better. Amongst the input WMs, we can observe a huge variation in the output of the RDPS model. High errors are concentrated across the region from north west to south east.

Temporally, the poorest performance was observed for GEFS and GEPS as shown in figure 11. The *best* performance is from the input WM RDPS, which is followed by the trained ML models of FNN and CNN.

D. MaxE - MAXIMUM ERROR METRIC

Amongst the input WMs, the best performing one is REPS. The best performance comes from the baseline IMM with 8.65 mm/day, which is closely followed by FNN with 8.67 mm/day.

The maximum error metric produced a spatial spectrum that is similar across the models, both input and trained as shown in figure 12. The high-error zones are concentrated around the central eastern, south eastern and the north-western parts of the map. The worst performing input WM is NAM at 13.18 and the best one is REPS at 9.16. For this metric, every single trained ML model improved over the input WMs, reducing the spatial mean of the error below 9 (with the exception of RFR). The best trained ML model is FNN at 8.67 which is followed by CNN at 8.71. We do note that there is a visible correlation with the observed precipitation, showing that high errors come from wet zones and the low errors come from dry zones.

In the temporal spectrum shown in figure 13, it can be observed that RDPS and NAM show a higher error value on

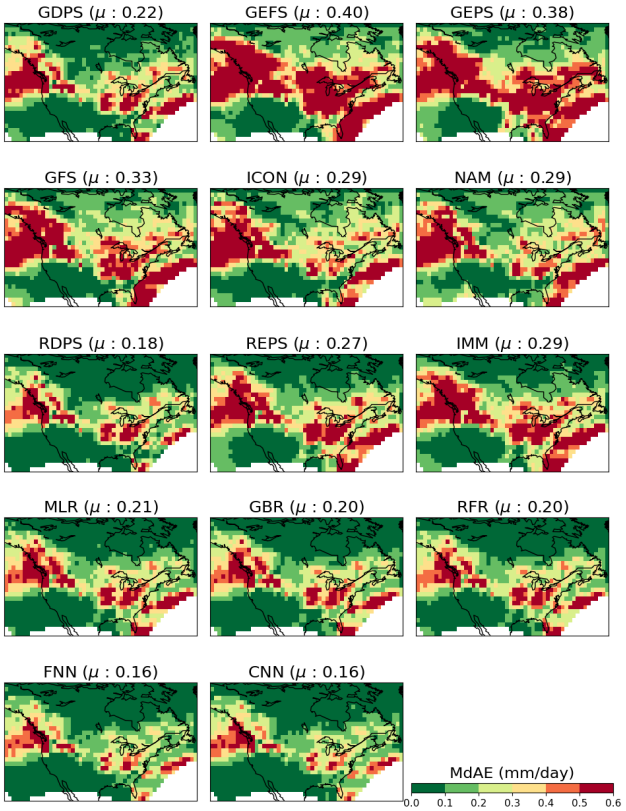


FIGURE 10. Median absolute error (spatial). Lower is better.

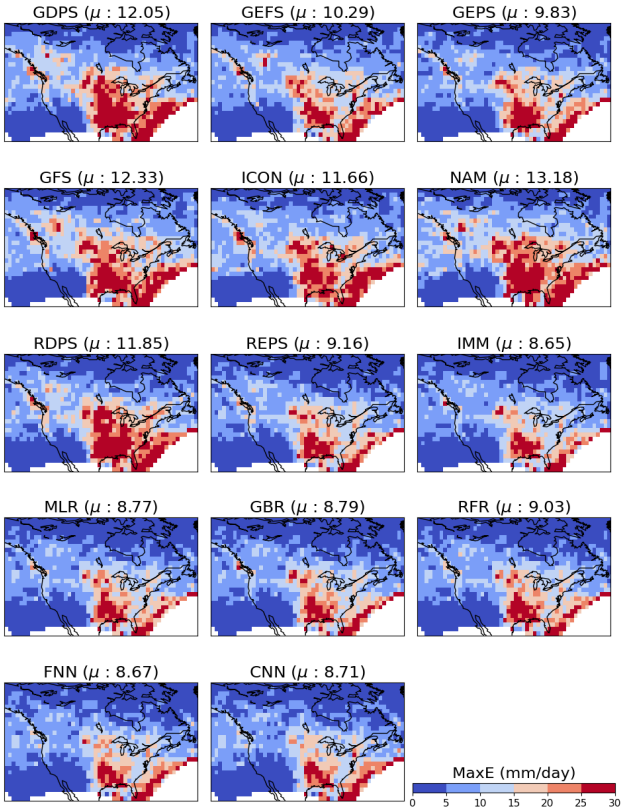


FIGURE 12. Maximum error (spatial). Lower is better.

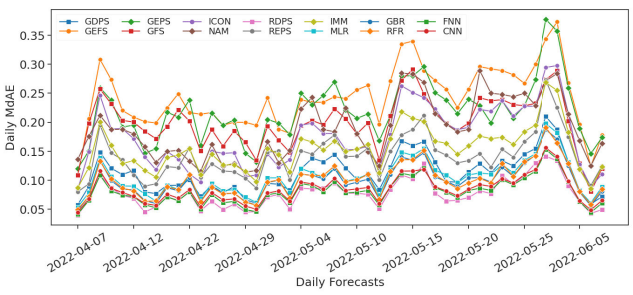


FIGURE 11. Median absolute error (temporal). Lower is better.

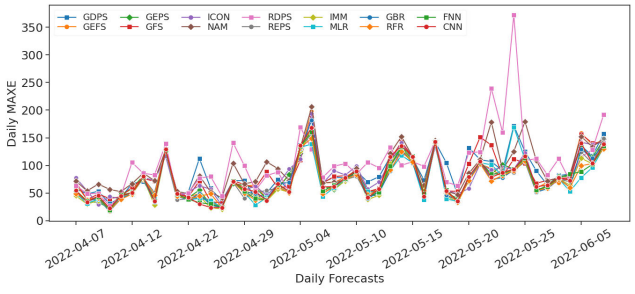


FIGURE 13. Maximum error (temporal). Lower is better.

most days, and the other input WMs exhibit fairly compact range of errors. The trained ML models provide the lowest error across the daily forecast dates.

E. CC - PEARSON CORRELATION COEFFICIENT METRIC

In the spatial context, this metric results in significant variation amongst the input WMs shown in figure 14. On the lowest end, we see a mean value of 0.72 for NAM. The highest correlation coefficient was observed for GEPS at 0.83, with the other input WMs yielding correlation coefficient values in the high 0.70's and low 0.80's. For this metric, all the trained ML models show superior performance with relatively low cross-model variation, ranging from 0.86 to 0.88. It appears the lowest correlation is observed at the southern end whereas the best results appear at the west and the north east of the map.

F. RB - RELATIVE BIAS METRIC

Figure 16 illustrates the spatial distribution of the relative biases, given the models. Noticeable variances are present across the input WMs. Over our particular validation set, all input WMs have positive relative bias. Magnitude-wise, the poorest performance comes from NAM with 1.56. The other input WMs perform significantly better. Decidedly, for this dataset, the trained neural network models are the best performing ones with values slightly below zero.

Temporal investigation reveals a similar result (figure 17). Again, trained neural networks (FNN and CNN) perform significantly better than the rest of the models.

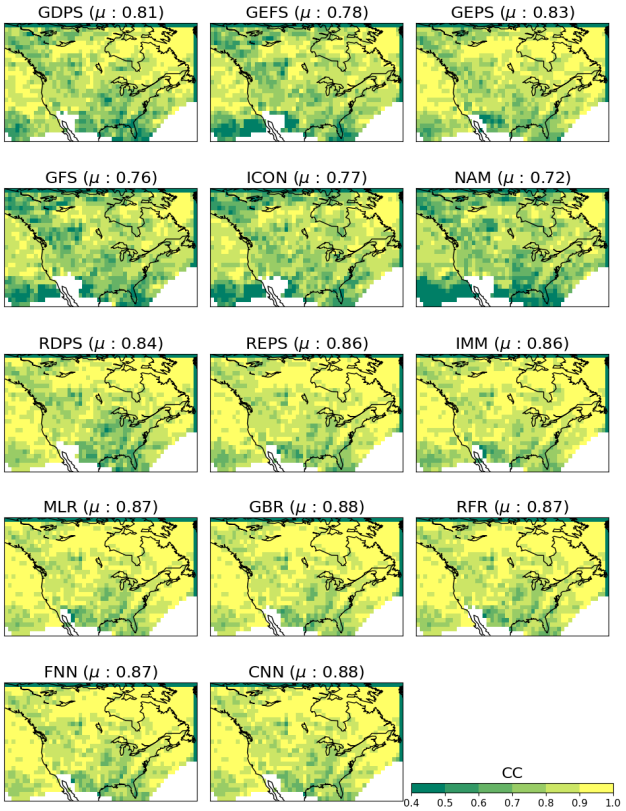


FIGURE 14. Correlation coefficient (spatial). Higher is better.

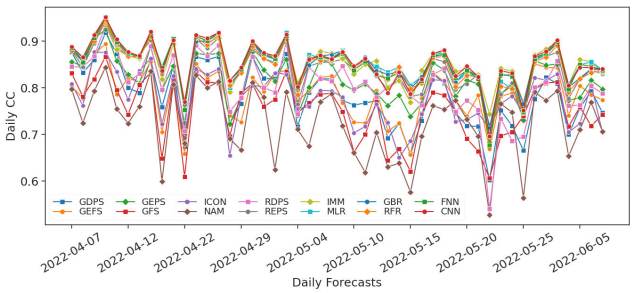


FIGURE 15. Correlation coefficient (temporal). Higher is better.

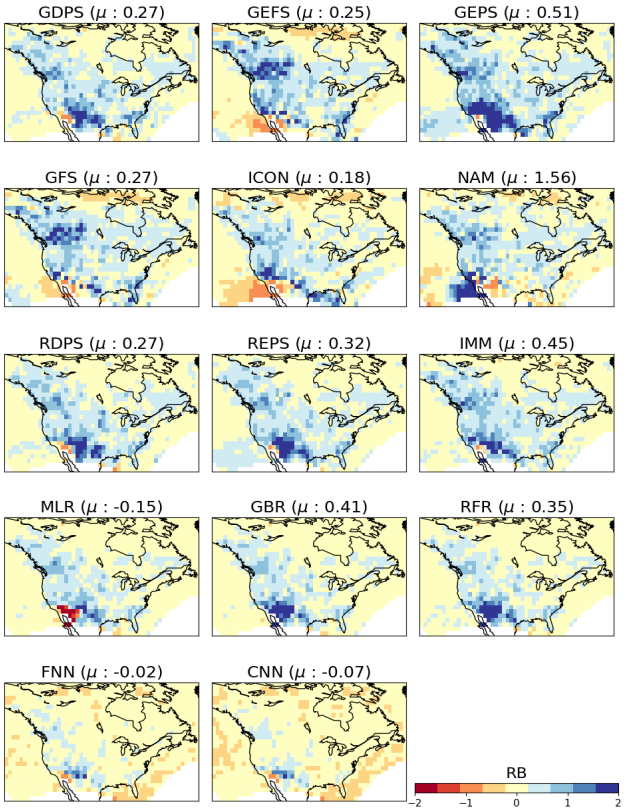


FIGURE 16. Relative bias (spatial). The closer to zero, the better.

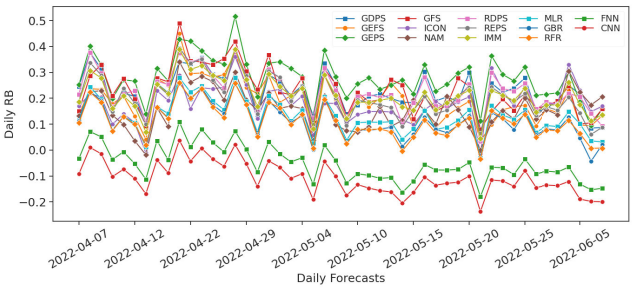


FIGURE 17. Relative bias (temporal). The closer to zero, the better.

G. POD - PROBABILITY OF DETECTION METRIC

For the probability of detection metric, in the spatial context, the best result among all was delivered by the input WM GEPS at 97%. It is significantly ahead of the remaining input WMs which range from 90% to 76%. As for the trained ML models, we see a more consistent performance pattern ranging from 95% to 93%. For this metric, the neural networks lag behind the other trained ML models, albeit by a small margin. Notably, the POD is lower around the drier climates, which is in line with the above-zero relative bias shown by the input WMs. These are shown in figure 18.

Figure 19 shows the temporal results. In temporal terms, GEPS once again achieved the highest result, with REPS as the second best. These WMs are followed by the baseline IMM and the remaining trained ML models. The significant difference in terms of the POD metric values between FNN and the rest of the models is also noteworthy.

H. FAR - FALSE ALARM RATIO METRIC

The spatial performance for this metric is illustrated in figure 20. For the false alarm ratio, spatially, a cluster of high error regions are observed around southern United States and the Gulf of Mexico and the Canadian Prairies for all the models.

Specifically, for GEPS, the high FAR value is due to the fact that the aforementioned error regions are also the largest.

As a result, this model sustained the highest false alarm ratio across the spatial spectrum, with NAM, REPS and ICON models resulting in lower FAR values. The best performing model, is the input WM RDPS, which is notable because it is

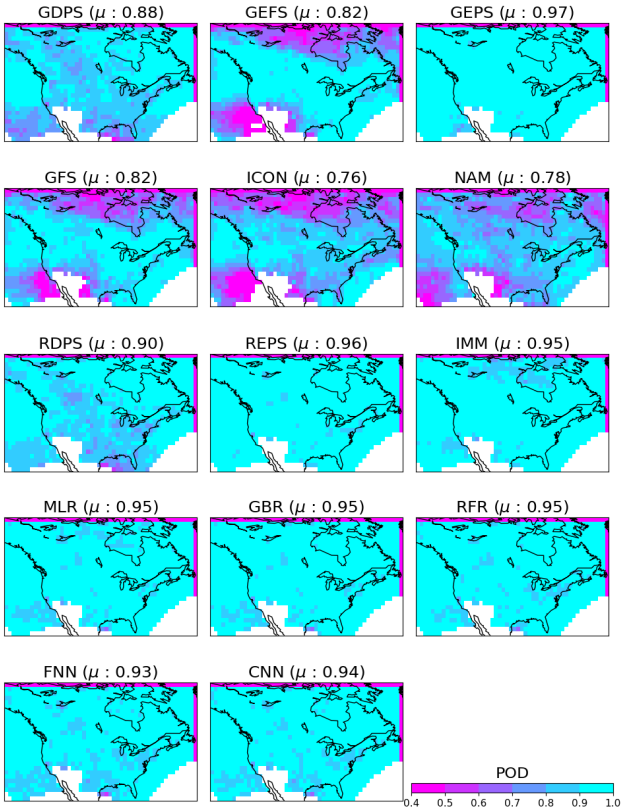


FIGURE 18. Probability of detection (spatial). Higher is better.

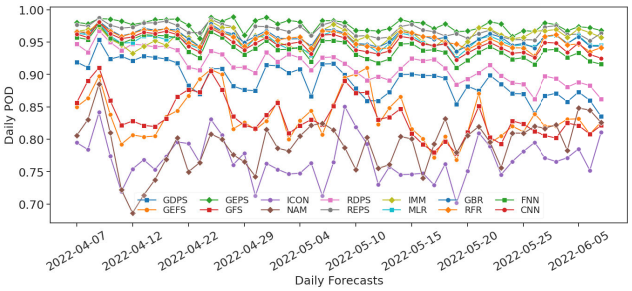


FIGURE 19. Probability of detection (temporal). Higher is better.

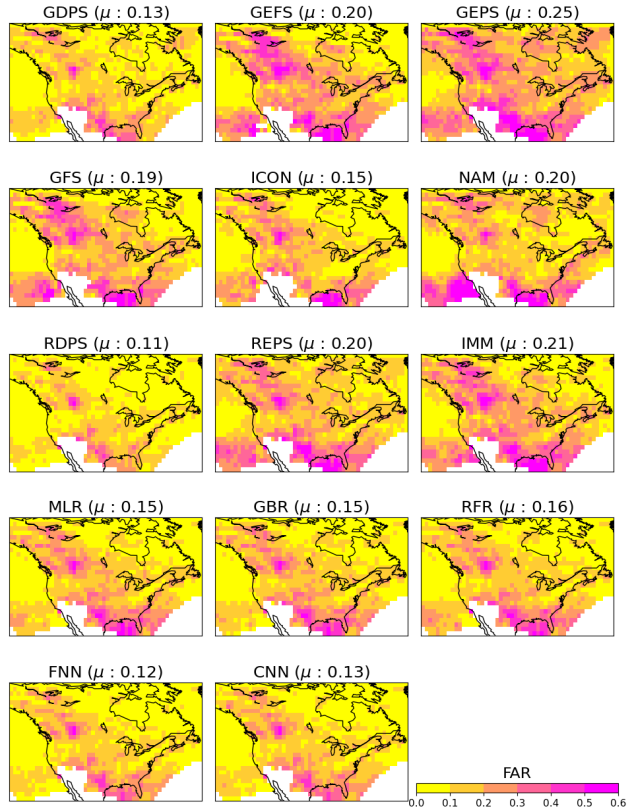


FIGURE 20. False alarm ratio (spatial). Lower is better.

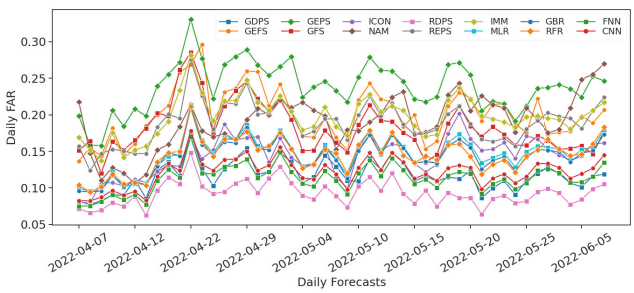


FIGURE 21. False alarm ratio (temporal). Lower is better.

also performed fairly well in terms of the POD metric (0.9). The metric value for the trained ML models is essentially between the RDPS and the other input WMs. The worst performance is observed by baseline IMM, which is poorer than 7 of 8 input WMs. The performance of FNN and CNN were slightly worse than REPS but they were still comparable.

Figure 21 shows the temporal performance. We see RDPS figures at the bottom of the chart by a comfortable margin, followed by GDPS, FNN and CNN. The poor performance of GEPS and GEFS are once again visible.

I. CSI - CRITICAL SUCCESS INDEX METRIC

Spatially, CNN and FNN are the best performing trained ML models. 4 of the 8 input WMs performed badly. These are GEFS, GFS, ICON, and NAM. The CSI values for the aforementioned models are all below 0.70. GEPS, GDPS, and REPS performed moderately well with the GEPS model at 0.73 and the GDPS and REPS at 0.78. RDPS performed the best at 0.82. For the trained ML models, we see on average, better performance than most of the input WMs. The baseline IMM displayed a modest performance at 0.76. The CSI values for the trained ML models are above 0.80. The overall best results are from the neural network models at 0.83. The results can be seen in figure 22.

Temporally, we see three clusters in figure 23. NAM, GEFS, ICON, GFS can be categorized as poor performing WMs. GEPS, baseline IMM, REPS and GDPS can be categorized as moderately performing ones. The remaining models (NWP and trained) achieved higher critical score

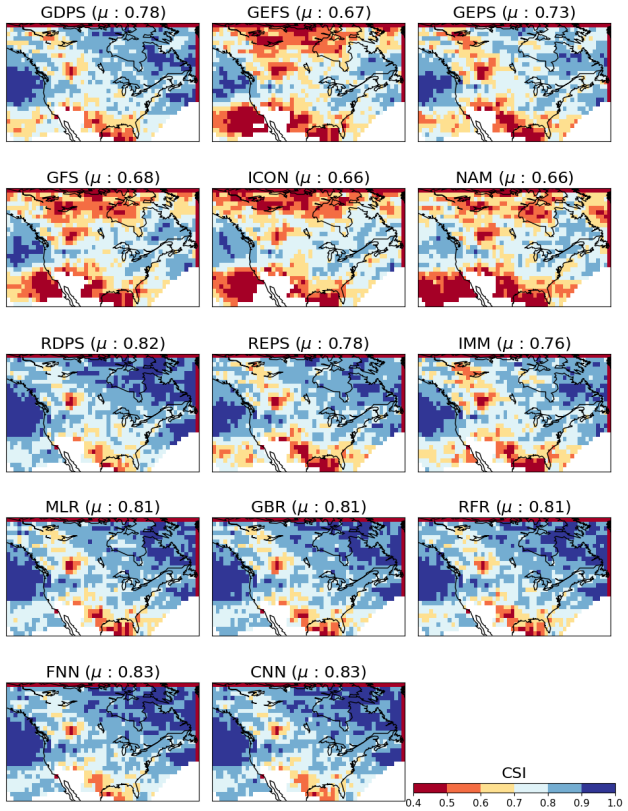


FIGURE 22. Critical score index (spatial). Higher is better.

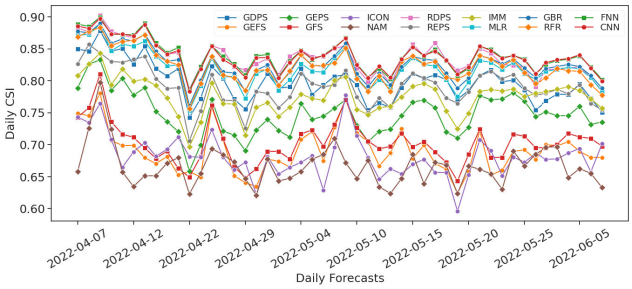


FIGURE 23. Critical score index (temporal). Higher is better.

index values. Near the very top of the list, we see that RDPS, FNN and CNN models perform well, similar to the spatial spectrum.

V. OVERALL RESULTS AND DISCUSSION

In this section, we present the overall performance of the five trained ML models as well as the input WMs in terms of the terms of the 9 evaluation metrics used in this study. Table 7 we give the results for the entire spatio-temporal test set for all the metrics.

In terms of the MAE metric, we see that REPS is the best performing input WM with the value of 0.96. It is followed by RDPS. The worst performer is NAM. For the trained ML models, the worst performer is the baseline IMM and the best performers are CNN and FNN with the respective values of 0.78 and 0.79. It is noteworthy that the performance of all trained ML models surpassed all the input NWP forecasts, suggesting that the application of machine learning techniques to improve precipitation forecasting is promising.

In terms of the RMSE metric, a similar set of outcomes are observed. The worst performer is NAM at 3.81 whereas the best performer is REPS among the input WMs. Once again, the trained ML models outperform the input WMs, with RMSE score ranging from 2.61 to 2.53. For this metric, the best performer is FNN, which is closely followed by CNN.

In terms of the MdAE metric, the performance gain is less evident. In fact, the overall best performer is the input WM RDPS at 0.07, which is followed by the neural network models with 0.08. Still, overall, the trained ML models perform better. The worst performance was by GEFS.

In terms of the MaxE metric, the best performance comes from the trained RFR (random forest regressor). It is notably better than the other models. It is then followed by the input WM REPS.

We do see higher correlation coefficient values for the trained ML models. There is relatively small variation. In this case RFR performed slightly worse than the others. The best performing input WM is REPS with NAM performing the worst.

We see the lowest relative bias from the neural network models. Overall, the trained ML models performed better than the input WMs except for the baseline IMM, which was mediocre. The worst bias was observed for GEPS.

In general, all trained ML models resulted in reasonable POD values. GEPS and REPS input WMs performed the best in terms of the probability of detection metric.

For the false alarm ratio metric, the best performance comes from FNN and RDPS models. They are followed by the CNN models. The poorest performance comes from GEPS, and together with the POD metric, it suggests that it has a bias towards regions that are wet. Overall, the trained ML models offer a better set of results, except for the baseline IMM which was relatively poor for this metric.

For the CSI metric, we see the best performance was demonstrated by FNN and CNN. They improve on the most input NWP forecasts, except for RDPS. The worst performer is NAM.

When all the metrics are considered, we see that the machine learning approaches do improve over the individual input NWP forecasts. Especially the neural network approaches seem to add noticeable value, showing superior performance across most metrics. For the individual WMs, RDPS offered the best performance in terms of most of the presented metrics. On the other hand, it performed significantly worse in terms of the RMS metric and MAE metric. This is also where the neural network models showed the most gains.

It is also noteworthy that considering its trivial composition, the performance improvement provided by baseline Input Means Model (IMM) is quite significant. It involves

TABLE 7. Results for the overall metrics for the entire spatio-temporal set. Best values are highlighted. The top 8 rows are input WMs. IMM is the baseline model.

	MAE	RMSE	MdAE	MaxE	CC	RB	POD	FAR	CSI
GDPS	1.15	3.57	0.11	172	0.78	0.22	0.89	0.15	0.77
GEFS	1.18	3.05	0.24	198	0.79	0.18	0.85	0.24	0.67
GEPS	1.12	2.97	0.21	188	0.81	0.27	**0.97**	0.27	0.71
GFS	1.28	3.53	0.19	191	0.75	0.21	0.82	0.21	0.68
ICON	1.21	3.40	0.17	193	0.78	0.17	0.78	0.17	0.67
NAM	1.31	3.81	0.19	206	0.73	0.15	0.77	0.20	0.64
RDPS	1.04	3.52	**0.07**	373	0.80	0.21	0.91	**0.13**	**0.81**
REPS	0.96	2.76	0.14	157	0.84	0.19	0.96	0.21	0.76
IMM	0.95	2.61	0.15	168	**0.85**	0.20	0.94	0.22	0.75
MLR	0.86	2.61	0.10	169	**0.85**	0.12	0.94	0.17	0.79
GBR	0.85	2.61	0.10	182	**0.85**	0.10	0.95	0.17	0.80
RFR	0.87	2.67	0.10	**149**	0.84	0.10	0.95	0.17	0.79
FNN	0.79	**2.53**	0.08	160	**0.85**	**-0.06**	0.93	**0.13**	**0.81**
CNN	**0.78**	2.54	0.08	169	**0.85**	-0.12	0.94	0.14	**0.81**

no training and yet is able to outperform every model in terms of MAE and RMSE. Random Forest Regressor model improves upon baseline IMM, and achieves the best performance in terms of MaxE. Multiple Linear Regressor and Gradient Boosting Regressor models likewise improve upon the baseline IMM for the majority of the metrics. Yet, the neural networks models offered the most significant gains.

VI. CONCLUSION

In this study, we have explored the capabilities of 5 different machine learning techniques to produce precipitation forecasts. We collected and extracted sample precipitation data covering most of the continental USA and Canada. Extensive preprocessing and feature selection tasks were performed on the spatio-temporal dataset. We selected eight WMs as input for training six classical machine learning models. Experiments show that machine learning approaches can improve weather forecast predictions in terms of 9 different metrics considered, and the best performers are the neural networks.

Future work includes considering additional input WMs, collection of a larger dataset over a longer span of time, and investigation of alternate deep neural network frameworks such as Generative Adversarial Networks, graph neural networks. A comparative performance analysis for longer range predictions is also a potential research area. The precipitation forecasts in this study were produced for daily (24-h) periods. While these are useful, many input WMs provide forecasts at increments as small as 1-h. To improve the usefulness of these forecasts, smaller temporal scales could be studied in the future. In addition, our forecasts were produced on a relatively coarse grid (0.125 degrees), which is not suitable for resolving smaller-scale phenomena such as convective storms. Using input models which are available at higher spatial and temporal resolutions may increase the utility of the resulting forecasts.

VII. CONFLICT OF INTEREST STATEMENT

There is no conflict of interest with the funders.

REFERENCES

[1] Z. Pu and E. Kalnay, *Numerical Weather Prediction Basics: Models, Numerical Methods, and Data Assimilation*. Berlin, Germany: Springer, 2019, pp. 67–97.

[2] A. Y. Hou, "The global precipitation measurement mission," *Bull. Amer. Meteorol. Soc.*, vol. 95, pp. 701–722, May 2014.

[3] T. M. Hamill, E. Engle, D. Myrick, M. Peroutka, C. Finan, and M. Scheuerer, "The U.S. national blend of models for statistical post-processing of probability of precipitation and deterministic precipitation amount," *Monthly Weather Rev.*, vol. 145, no. 9, pp. 3441–3463, Sep. 2017.

[4] T. M. Hamill, D. R. Stovern, and L. L. Smith, "Improving national blend of models probabilistic precipitation forecasts using long time series of reforecasts and precipitation reanalyses. Part I: Methods," *Monthly Weather Rev.*, vol. 151, no. 6, pp. 1521–1534, Jun. 2023.

[5] D. R. Stovern, T. M. Hamill, and L. L. Smith, "Improving national blend of models probabilistic precipitation forecasts using long time series of reforecasts and precipitation reanalyses. Part II: Results," *Monthly Weather Rev.*, vol. 151, no. 6, pp. 1535–1550, Jun. 2023.

[6] J. P. Craven, D. E. Rudack, and P. E. Shafer, "National blend of models: A statistically post-processed multi-model ensemble," *J. Oper. Meteorol.*, vol. 8, no. 1, pp. 1–14, Jan. 2020.

[7] R. Buizza and T. N. Palmer, "Impact of ensemble size on ensemble prediction," *Monthly Weather Rev.*, vol. 126, no. 9, pp. 2503–2518, Sep. 1998.

[8] W. Li, Q. Duan, Q. J. Wang, S. Huang, and S. Liu, "Evaluation and statistical post-processing of two precipitation reforecast products during summer in the Mainland of China," *J. Geophys. Res., Atmos.*, vol. 127, no. 12, Jun. 2022, Art. no. e2022JD036606.

[9] S. K. Jha, D. L. Shrestha, T. A. Stadnyk, and P. Coulibaly, "Evaluation of ensemble precipitation forecasts generated through post-processing in a Canadian catchment," *Hydrol. Earth Syst. Sci.*, vol. 22, no. 3, pp. 1957–1969, Mar. 2018.

[10] D. E. Robertson, D. L. Shrestha, and Q. J. Wang, "Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting," *Hydrol. Earth Syst. Sci.*, vol. 17, no. 9, pp. 3587–3603, Sep. 2013.

[11] E. E. Ebert, "Ability of a poor man's ensemble to predict the probability and distribution of precipitation," *Monthly Weather Rev.*, vol. 129, no. 10, pp. 2461–2480, Oct. 2001.

[12] M. Min, C. Bai, J. Guo, F. Sun, C. Liu, F. Wang, H. Xu, S. Tang, B. Li, D. Di, L. Dong, and J. Li, "Estimating summertime precipitation from Himawari-8 and global forecast system based on machine learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 5, pp. 2557–2570, May 2019.

[13] P. T. Nastos, K. P. Moustris, I. K. Larissi, and A. G. Paliatsos, "Rain intensity forecast using artificial neural networks in Athens, Greece," *Atmos. Res.*, vol. 119, pp. 153–160, Jan. 2013.

[14] P. Joshi, M. S. Shekhar, A. Kumar, and J. K. Quamara, "Artificial neural network model for precipitation forecast over western Himalaya using satellite images," *MAUSAM*, vol. 73, no. 1, pp. 83–90, Jan. 2022.

[15] C. K. Sønderby, L. Espeholt, J. Heek, M. Dehghani, A. Oliver, T. Salimans, S. Agrawal, J. Hickey, and N. Kalchbrenner, "MetNet: A neural weather model for precipitation forecasting," 2020, *arXiv:2003.12140*.

[16] A. Y. Barrera-Animas, L. O. Oyedele, M. Bilal, T. D. Akinosho, J. M. D. Delgado, and L. A. Akanbi, "Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting," *Mach. Learn. Appl.*, vol. 7, Mar. 2022, Art. no. 100204. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S266682702100102X

[17] B. Bochenek and Z. Ustrnul, "Machine learning in weather prediction and climate analyses—Applications and perspectives," *Atmosphere*, vol. 13, no. 2, p. 180, Jan. 2022.

[18] X. Ren, X. Li, K. Ren, J. Song, Z. Xu, K. Deng, and X. Wang, "Deep learning-based weather prediction: A survey," *Big Data Res.*, vol. 23, Feb. 2021, Art. no. 100178.

[19] D. N. Tuyen, T. M. Tuan, X.-H. Le, N. T. Tung, T. K. Chau, P. Van Hai, V. C. Gerogiannis, and L. H. Son, "RainPredRNN: A new approach for precipitation nowcasting with weather radar echo images based on deep learning," *Axioms*, vol. 11, no. 3, p. 107, Feb. 2022. [Online]. Available: https://www.mdpi.com/2075-1680/11/3/107

[20] Y. Wang, H. Wu, J. Zhang, Z. Gao, J. Wang, P. S. Yu, and M. Long, "PredRNN: A recurrent neural network for spatiotemporal predictive learning," 2021, *arXiv:2103.09504*.

[21] A. Haidar and B. Verma, "Monthly rainfall forecasting using one-dimensional deep convolutional neural network," *IEEE Access*, vol. 6, pp. 69053–69063, 2018.

[22] S. Yao, H. Chen, E. J. Thompson, and R. Cifelli, "An improved deep learning model for high-impact weather nowcasting," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 7400–7413, 2022.

[23] S. Ravuri, K. Lenc, M. Willson, D. Kangin, R. Lam, P. Mirowski, M. Fitzsimons, M. Athanassiadou, S. Kashem, S. Madge, R. Prudden, A. Mandhane, A. Clark, A. Brock, K. Simonyan, R. Hadsell, N. Robinson, E. Clancy, A. Arribas, and S. Mohamed, "Skillful precipitation nowcasting using deep generative models of radar," 2021, *arXiv:2104.00954*.

[24] P. Hess and N. Boers, "Deep learning for improving numerical weather prediction of heavy rainfall," *J. Adv. Model. Earth Syst.*, vol. 14, no. 3, Mar. 2022, Art. no. e2021MS002765.

[25] D. Cho, C. Yoo, B. Son, J. Im, D. Yoon, and D.-H. Cha, "A novel ensemble learning for post-processing of NWP model's next-day maximum air temperature forecast in summer using deep learning and statistical approaches," *Weather Climate Extremes*, vol. 35, Mar. 2022, Art. no. 100410.

[26] M. C. V. Ramírez, H. F. de Campos Velho, and N. J. Ferreira, "Artificial neural network technique for rainfall forecasting applied to the São Paulo region," *J. Hydrol.*, vol. 301, nos. 1–4, pp. 146–162, Jan. 2005.

[27] D. Gagne, A. Mcgovern, and M. Xue, "Machine learning enhancement of storm-scale ensemble probabilistic quantitative precipitation forecasts," *Weather Forecasting*, vol. 29, pp. 1024–1043, Aug. 2014.

[28] Z. Fan, W. Li, Q. Jiang, W. Sun, J. Wen, and J. Gao, "A comparative study of four merging approaches for regional precipitation estimation," *IEEE Access*, vol. 9, pp. 33625–33637, 2021.

[29] J. Frnda, M. Durica, J. Rozhon, M. Vojtekova, J. Nedoma, and R. Martinek, "ECMWF short-term prediction accuracy improvement by deep learning," *Sci. Rep.*, vol. 12, no. 1, pp. 1–11, May 2022.

[30] K. Zhou, J. Sun, Y. Zheng, and Y. Zhang, "Quantitative precipitation forecast experiment based on basic NWP variables using deep learning," *Adv. Atmos. Sci.*, vol. 39, pp. 1472–1486, Apr. 2022.

[31] C.-M. Ko, Y. Y. Jeong, Y.-M. Lee, and B.-S. Kim, "The development of a quantitative precipitation forecast correction technique based on machine learning for hydrological applications," *Atmosphere*, vol. 11, no. 1, p. 111, Jan. 2020. [Online]. Available: https://www.mdpi.com/2073-4433/11/1/111

[32] L. Xu, N. Chen, X. Zhang, Z. Chen, C. Hu, and C. Wang, "Improving the North American multi-model ensemble (NMME) precipitation forecasts at local areas using wavelet and machine learning," *Climate Dyn.*, vol. 53, nos. 1–2, pp. 601–615, Jul. 2019.

[33] V. M. Krasnopolsky and Y. Lin, "A neural network nonlinear multimodel ensemble to improve precipitation forecasts over continental U.S.," *Adv. Meteorol.*, vol. 2012, pp. 1–11, Jan. 2012.

[34] L. Espeholt, S. Agrawal, C. Sønderby, M. Kumar, J. Heek, C. Bromberg, C. Gazen, J. Hickey, A. Bell, and N. Kalchbrenner, "Skillful twelve hour precipitation forecasts using large context neural networks," 2021, *arXiv:2111.07470*.

[35] Y. Fan, V. Krasnopolsky, H. van den Dool, C.-Y. Wu, and J. Gottschalck, "Using artificial neural networks to improve CFS week-3–4 precipitation and 2-m air temperature forecasts," *Weather Forecasting*, vol. 38, no. 5, pp. 637–654, May 2021.

[36] N. Gasset, T. Milewski, A. Beaulne, F. Dupont, and A. Zadra. *The Global Deterministic Prediction System (GDPS) Version 8.0.0 of the Meteorological Service (MSC) of Canada*. Environment and Climate Canada. Accessed: Jul. 7, 2023. [Online]. Available: https://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/tech_specifications/tech_specifications_GDPS_e.pdf

[37] National Oceanic and Atmospheric Administration. *Global Ensemble Forecast System*. Accessed: Jul. 7, 2023. [Online]. Available: https://www.emc.ncep.noaa.gov/emc/pages/numerical_forecast_systems/gefs.php

[38] X. Deng, J. S. Fontecilla, and P. Houtekamer. *The Global Ensemble Prediction System (GEPS) Version 7.0 of the Meteorological Service (MSC) of Canada*. Environment and Climate Canada. Accessed: Jul. 7, 2023. [Online]. Available: https://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/tech_specifications/tech_specifications_GEPS_e.pdf

[39] National Oceanic and Atmospheric Administration. *Gfs*. Accessed: Jul. 7, 2023. [Online]. Available: https://www.emc.ncep.noaa.gov/emc/pages/numerical_forecast_systems/gfs.php

[40] Deutscher Wetterdienst. *Icon (Icosahedral Nonhydrostatic) Model*. Accessed: Jul. 7, 2023. [Online]. Available: https://www.dwd.de/EN/research/weatherforecasting/num_modelling/01_num_weather_prediction_modells/icon_description.html

[41] National Oceanic and Atmospheric Administration. *North American Mesoscale*. Accessed: Jul. 7, 2023. [Online]. Available: https://www.emc.ncep.noaa.gov/emc/pages/numerical_forecast_systems/nam.php

[42] T. Milewski, A. Zadra, and J.-F. Caron. *The Regional Deterministic Prediction System (RDPS) Version 8.0.0 of the Meteorological Service of Canada (MSC)*. Environment and Climate Canada. Accessed: Jul. 7, 2023. [Online]. Available: https://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/tech_specifications/tech_specifications_RDPS_e.pdf

[43] *The Regional Ensemble Prediction System (REPS) Version 4.0 of the Meteorological Service of Canada (MSC) Technical Specifications*. Environment and Climate Canada. Accessed Jul. 7, 2023. [Online]. Available: https://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/tech_specifications/tech_specifications_REPS_e.pdf

[44] F. Lespinas. *Regional Deterministic Precipitation Analysis System (CAPA-RDPA), Implementation Of Version 5.2.0, Technical Note*. Environment and Climate Canada. Accessed: Jul. 7, 2023. [Online]. Available: https://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/lib/technote_capa_rdpa_e.pdf

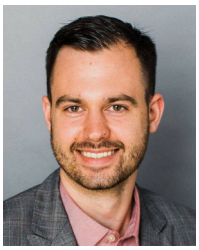
[45] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001. [Online]. Available: http://www.jstor.org/stable/2699986

[46] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[47] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature," *Geoscientific Model Develop.*, vol. 7, no. 3, pp. 1247–1250, Jun. 2014.

CENKER SENGOZ received the B.Sc. degree in computer engineering from Bilkent University, Turkey, and the M.Sc. degree in applied computer science from The University of Winnipeg, Canada. He has industrial experience in software development, database management, and embedded systems. He has been collaborating with The University of Winnipeg, to conduct research on machine learning applications. His research interests include computer vision, deep neural networks, natural language processing, and the theory and applications of set approximation. The research and development that he contributed in industry and the academia, found applications in road condition detection, weather prediction, crowd counting, face recognition, image segmentation, and web information labeling.

SHEELA RAMANNA received the B.S. degree in EE and the M.S. degree in CS from Osmania University, India, and the Ph.D. degree in CS from Kansas State University, USA. She is a Professor and the past Chair of the Applied Computer Science Department, The University of Winnipeg (UW). She is the Co-Founder of the UW ACS Graduate Studies Program. Her research was funded by NSERC Discovery/Engage/Alliance/MITACS Accelerate grants. She has published over 55 articles in the past ten years and has been supervising award winning graduate students, since 2012. She has co-edited a book *Emerging Paradigms in Machine Learning* (ML) (Springer, 2013). Her current research is on foundations and applications of computational intelligence (CI) methods (rough, near, and rough-fuzzy sets), with applications in natural language processing and multimodal information processing. Her research interests also include topological data analysis and persistent homology-based machine learning. She was a recipient of the 2015 TUBITAK Fellowship, Turkey. She has served as the Program Co-Chair for IJCRS2021, MIWAI2013, RSKT2011, RSCTC2010, and JRS2007. She serves on the EB for *Transactions on Rough Sets* (TRS) (Springer) and *EAAI*, in June 2022; and on the Advisory Board for the *International Journal of Rough Sets and Data Analysis*. She is a Managing Editor of *TRS* and a Senior Member of the International Rough Set Society.



SCOTT KEHLER received the B.Sc. and M.Sc. degrees from the University of Manitoba, Winnipeg, MB, Canada. He is the President and the Chief Scientist with Weatherlogics Inc. His work involves developing novel ways to improve weather prediction.

RUSHIL GOOMER received the B.Tech. degree in computer science from Bennett University, India. He is currently pursuing the master's degree in applied computer science and society with The University of Winnipeg, with a focus on enhancing precipitation forecasting of numerical weather prediction (NWP) models. He has gained practical experience in the field through his involvement with the Central Scientific Instruments Organization (CSIO), where he provided technology consulting and applied data science techniques to develop water quality monitoring systems. He is dedicated to exploring and implementing innovative solutions in the realm of machine learning and positively contribute to the society.

PAUL PRIES received the B.Comp. degree from Queens University, Kingston, ON, Canada. He is a scientific software developer. His work has involved the development of software to process meteorological data.