Improving Precipitation Forecasts with Convolutional Neural Networks

Anirudhan Badrinath[1], Luca Delle Monache[2], Negin Hayatbini[3], William Eric Chapman[4], Forest Cannon[4], and F. Martin Ralph[5]

[1]University of California, Berkeley
[2]University of California San Diego
[3]Center for Western Weather and Water Extremes, Scripps Institution of Oceanography
[4]University of California, San Diego
[5]SIO

November 21, 2022

Abstract

Traditional post-processing methods have relied on point-based applications that are unable to capture complex spatial precipitation error patterns. With novel ML methods using convolution to more effectively identify and reduce spatial biases, we propose a modified U-Net convolutional neural network (CNN) to post-process daily accumulated precipitation over the US west coast. For training, we leverage 34 years of deterministic Western Weather Research and Forecasting (West-WRF) reforecasts. On an unseen 4-year data set, the trained CNN yields a 12.9-15.9% reduction in root mean-square error (RMSE) over West-WRF for lead times of 1-4 days. Compared to an adapted Model Output Statistics baseline, the CNN reduced RMSE by 7.4-8.9% for all events. Effectively, the CNN adds more than a day of predictive skill when compared to West-WRF. The CNN outperforms the other methods also for the prediction of extreme events, highlighting a promising path forward for improving precipitation forecasts.

Improving Precipitation Forecasts with Convolutional Neural Networks

Anirudhan Badrinath[1], Luca Delle Monache[2], Negin Hayatbini[2], Will Chapman[2], Forest Cannon[2], Marty Ralph[2]

[1]Department of Computer Science, University of California Berkeley, Berkeley, California, United States
[2]Center for Western Weather and Water Extremes, Scripps Institution of Oceanography, University of California San Diego, La Jolla, California, United States

Key Points:

- We adapted a U-Net convolutional neural network (CNN) architecture as a post-processing framework.
- The precipitation class imbalance was addressed by the dual ML model approach.
- The proposed method provides greater numerical accuracy over all lead times.

Corresponding author: Anirudhan Badrinath, `abadrinath@berkeley.edu`

Abstract

Traditional post-processing methods have relied on point-based applications that are unable to capture complex spatial precipitation error patterns. With novel ML methods using convolution to more effectively identify and reduce spatial biases, we propose a modified U-Net convolutional neural network (CNN) to post-process daily accumulated precipitation over the US west coast. For training, we leverage 34 years of deterministic Western Weather Research and Forecasting (West-WRF) reforecasts.

On an unseen 4-year data set, the trained CNN yields a 12.9-15.9% reduction in root mean-square error (RMSE) over West-WRF for lead times of 1-4 days. Compared to an adapted Model Output Statistics baseline, the CNN reduced RMSE by 7.4-8.9% for all events. Effectively, the CNN adds more than a day of predictive skill when compared to West-WRF. The CNN outperforms the other methods also for the prediction of extreme events, highlighting a promising path forward for improving precipitation forecasts.

Plain Language Summary

Machine learning methods are used for accurate large-scale prediction by learning patterns from a vast amount of data. We demonstrate the utility of a computer vision-based machine learning technique for improving precipitation forecasts. Extreme precipitation events and atmospheric rivers, which contain narrow bands of water vapor transport, can cause millions in damages. We show that there is a significant increase in predictive accuracy for daily accumulated precipitation using these machine learning methods, which could result in significant societal benefits.

1 Introduction

The precipitation associated with atmospheric rivers (ARs), "a long, narrow, and transient corridor of strong horizontal water vapor transport" (AMS, 2019), replenishes the water supply but can also result in flooding over the western United States. ARs cause median economic losses in the tens to hundreds of millions of dollars for AR4 and AR5 ARs based on the AR scale developed by (Ralph et al., 2019). Further, ARs have been identified as the primary source of hydrologic flooding in the western United States (Corringham et al., 2019). Accurate and reliable predictions of precipitation can help in minimizing losses attributable to ARs or other weather phenomena (e.g., cut-off lows, narrow cold-frontal rainbands, etc.) and in better managing the water supply in the western United States (ODonnell et al., 2020).

Numerical weather prediction (NWP) are based on dynamical models that are built on current state-of-the-science knowledge of key atmospheric physics and numerical procedure. However, NWP accuracy is affected by initial condition errors, numerical approximations, and incomplete understanding and representation of all the relevant physical processes (Delle Monache et al., 2013; Vannitsem & Ghil, 2017; Collins & Allen, 2002; Nicolis & Nicolis, 2007).

NWP post-processing methods are designed to correct for the aforementioned deficiencies by learning the characteristics of NWP errors from a historical data set to then try to anticipate today forecast biases. These include downscaling methods, Kalman filters, model output statistics (MOS), and machine learning methods such as neural network models, decision trees, and multilinear regression models (Louka et al., 2008; Glahn & Lowry, 1972). Historically, post-processing methods, including machine learning methods, have operated on a point-by-point basis (Rasp & Lerch, 2018). Recently, convolutional neural networks (CNNs) have been proposed to correct satellite retrievals (Tao et al., 2016). CNNs have been shown to be a powerful regression tool in the domains of image analysis (Krizhevsky et al., 2012). For NWP, recent work by W. Chapman et al. (2019) has highlighted the efficacy of a CNN as a NWP post-processing method for the prediction of integrated vapor transport (IVT). It was shown that the CNN-based prediction resulted in 9-17% improvements in RMSE as compared to other methods (W. Chapman et al., 2019).

Traditional point-by-point approaches have been shown to be effective in improving the raw estimates of dynamical models (Glahn & Lowry, 1972), and are particularly valuable for certain applications, e.g., renewable energy (Alessandrini et al., 2015; Cervone et al., 2017). However, since spatial interdependence is ignored, at times non-physical fields with statistical anomalies are produced (Vannitsem & Ghil, 2017). Further, the predominantly "no rain" data points in the precipitation field poses an issue for standard machine learning methods which rely on balanced classes of data. To mitigate these issues, in this study we explore the potential of a recently developed machine learning method to post-process accumulated precipitation forecasts: a U-Net CNNs architecture (Ronneberger et al., 2015). U-Net CNNs are a form of artificial neural network , which have been used for both classification and regression tasks primarily focused on spatial data in the field of biomedical imaging. As such, this CNN leverages spatial interdependence by construction. Further, we propose and test a dual ML model structure to rectify the class imbalance in the sparse precipitation data.

In Section 2 we introduce the data used in this study. Section 3 describes the methodology, including evaluation strategy and skill scores. The results are presented in section 4. Conclusions are provided in section 5, where the potential of U-Net CNN as a tool for weather forecasting and future research is discussed.

2 Data and Methodology

2.1 Observational Data

The observed precipitation data used in this study is the Parameter-elevation Relationships on Independent Slopes Model (PRISM) dataset (PRISM Climate Group, 2004), which is constructed using data from the Cooperative Observer Program (COOP) and Snowpack Telemetry (SNOTEL) networks, and a variety of smaller networks (Daly et al., 2008). PRISM provides estimates of accumulated 24 hour precipitation data over the last 40 years over the contiguous United States (CONUS) at a spatial resolution of 4 km. Here we focus on the western United States region comprising California and Nevada.

The PRISM dataset was chosen as ground truth in this study due to its accuracy, comparable spatial resolution to the model reforecast data, and length of record. PRISM uses a comprehensive linear precipitation–elevation correction scheme that applies weights based on location to nearby stations, proximity to coast, topographic facets, boundary layer conditions, surrounding terrain height, and other terrain features (Daly et al., 2008). PRISM has been shown to perform well in challenging complex terrain settings when tested against independent station data (Daly et al., 2017). It has also been shown to produce reliably similar estimates of precipitation extremes when compared to other national in-situ based gridded datasets, while performing notably better than various reanalysis products (Gibson et al., 2019).

2.2 Model Reforecast Data

The NWP reforecast data which is being post-processed, was developed at the Center for Western Weather and Water Extremes. As input to the U-Net CNN, we use weather forecasts over a 3-km domain (Figure 1) of 34 water years (1985 to 2019) of the Western Weather Research and Forecasting (West-WRF) regional model (Martin et al., 2019) covering the western United States, California and Nevada. The forecasts are driven by initial and boundary conditions from the Global Forecasting System. The West-WRF regional model has shown forecast skill with a low intensity error for IVT and reduced dry and wet biases for precipitation over lead times from 1 to 7 days (see Steinhoff et al. (2020) for additional details on the reforecast).

To align the forecast spatially with the observation set, we regrid the forecasts with a nearest-neighbor approach to a 4-km resolution, to retain existing precipitation patterns and preserve global precipitation means. For temporal alignment with the observations, and given that the forecast are initialized daily at 0000 UTC, we calculate the accumulated daily precipitation offset by 12 hours to account for model spin-up. In other words, data from 12-h to 36-h after initialization of each West-WRF forecast is labelled as Day 1 forecast and is aligned with PRISM ground truth data.

2.3 Machine Learning Approach

The proposed CNN for forecast post-processing uses the U-Net architecture as a baseline, named after its distinctive U-shape model diagram (Ronneberger et al., 2015). Historically, this type of CNN has been used for biomedical image segmentation, but its application with weather forecasts is promising given its strength in rectifying spatial biases through image segmentation (W. Chapman et al., 2019). The model architecture consists of two phases. In the first phase, the model performs data compression through repeated convolutional layers to learn spatial features. This is followed by an expanding phase in which the output image is reconstructed using the learned features. We modified the U-Net architecture as introduced by Ronneberger et al. (2015) in several ways as detailed below to adapt it to the task of improving the skill of precipitation forecasts. The model along with these modifications is referred to as the modified U-Net CNN from here onwards.

West-WRF model output variables are used as predictors in the CNN. In particular, to generate Day 1 predictions, the normalized 24-h accumulated precipitation, and the 6, 12 and 18-h forecasts of 5-m specific humidity and 2-m temperature since forecast initialization are used. Similarly, for greater lead times, we use the same predictors offset by the lead times. These predictors are used because they provide significant insight into the ground truth precipitation (Richardson, 1922). It was determined through validation that as the number of input parameters was increased beyond these predictors and time granularity (e.g., hourly instead of 6 hourly), the efficiency and accuracy of the model decreased (Anelli et al., 2019).

The loss function used for the modified U-Net CNN is an asymmetric adaptation of the mean-square error that penalizes underprediction more than overprediction. It was observed through preliminary tests that the U-Net CNN tended to systematically underpredict extreme precipitation events, hence we chose to correct this bias as follows. We assign a hyperparameter $w_s > 1$ that multiplicatively weights underpredicted values as described in Equation S2 in the supplemental information. The value of w_s is determined by minimizing loss on the validation data set, which is consistent with the procedure to determine all hyperparameters.

To combat a tendency for neural networks to predict small non-zero values of precipitation for every grid cell due to millions of additions in its numerical computations (for example, a "zero" value might be predicted as 0.001), we leverage binary masking, during model training, for precipitation prediction (Hayatbini et al., 2019). Binary masking is a classification technique that generates a rain vs. no rain map for all grid points given the same input as the main post-processing framework. We use the same model architecture for training this binary mask predictor as the main post-processing model except the predictions (the numerical precipitation value) are replaced by indicator functions of the precipitation (i.e., rain vs. no rain). We train this completely separately from the main post-processing model. In other words, instead of predicting the amount of precipitation, we predict the probability of non-zero precipitation at that grid point. Then, we use masking to remove any values in the main numerical precipitation prediction that were predicted as likely having zero rain with over 50% probability by our binary mask. The loss function used in this case is the cross-entropy loss, a standard loss used in this kind of classification problems (Hayatbini et al., 2019). Figure S1 summarizes the aforementioned structure, located in the supplemental information.

Further, we propose a dual ML model solution to class imbalance between the occurrence of extreme and moderate precipitation events. We will refer to this as the dual model approach. For extreme events, traditional machine learning-based baselines such as MOS tend to underestimate the upper tail of the distribution and overestimate the moderate case due to the relatively low probability of extreme values in the distribution. To address this issue, we create separate U-Net CNN models for the more extreme events as classified by mean forecast accumulated precipitation above 2.5 mm. This corresponds with roughly all events below the 20th percentile total accumulated precipitation, which was determined through validation as an effective separation to mitigate the class imbalance issue. For the remaining events, we train a separate U-Net model to preserve predictive capability for the moderate case. Through this, we accomplish a tailored model for both extreme and moderate precipitation. While there exist deep learning techniques that resolve class imbalances in a more formal way such as data augmentation, they rely on mutating the data (e.g., stretching or cropping), which may be less desirable for post-processing problems with a numerical output (Perez & Wang, 2017). This is because these techniques produce an augmented input, but the numerical output (ground truth) then needs to be augmented too. Hence, we don't perform this and instead assume that general mean and total precipitation over a region is roughly consistent in distribution over water years.

Parameter tuning for the learning rate, the number of filters per layer, and loss function weights is accomplished through validation. The optimal hyperparameters were close to their default values as provided in Keras, the used machine learning library (Chollet et al., 2015). The values and more detailed information regarding hyperparameter tuning are provided in the supplementary information.

2.4 Testing and Evaluation

The CNN is evaluated over a chosen test set of 4 water years, which were selected based on categorical El Niño/Southern Oscillation years. We use one El Niño year (1997), one La Niña year (2011), and two ENSO neutral years: one historically wet and one dry year (years 2016 and 2013, respectively). ENSO years have been shown to dramatically effect West Coast precipitation regimes through large scale pressure patterns which significantly alter precipitation predictability (W. E. Chapman et al., 2021; Kumar & Hoerling, 1998). We also select particularly wet (2016/2017) and dry (2013/2014) years in which ENSO is in a neutral state, representing California drought conditions and a surplus of precipitation, respectively, without tropical SST forcing. We choose these years in order to test the skill of our methods in varied climate regimes and on a variety of precipitation events. The rest serves as the training set. We use a testing process that most closely mimics a production system in which we train one CNN model over all possible years except a singular testing year and a validation year (the latter used to tune the hyperparameters); this is done for all years, so we train 4 dual ML models in total (8 in total), each of which is not trained on their corresponding test year. We refer to this as "one-shot" training.

Traditional machine learning and dynamical post-processing frameworks were compared to the proposed U-Net CNN to assure its predictive accuracy and reliability over the chosen test set. Further, they offered a baseline for the CNN's forecasting skill. A prediction based on climatology was used to ensure that the CNN is consistent and reliable. It was constructed by averaging 30 days worth of observation data prior to any particular testing day over all years preceding it. The second comparison was with the West-WRF dynamical model, which is used as the input to the machine learning method. As such, any rectification of spatial or temporal biases over the West-WRF model would be directly reflected in the CNN's accuracy and errors. Further, we implemented a MOS based on a L1-regularized multilinear regression (Tibshirani, 1996). The MOS presents a more traditional ML framework that can be used as a baseline to the CNN. Similar to many other ML frameworks, the MOS leverages point-based learning as opposed to the strategy adopted in a CNN. Note that the multilinear regression is configured to use the same predictors (precipitation, humidity, temperature) as the CNN and uses the same "one-shot" training for consistency.

We evaluated the model using the following metrics: root-mean square error (RMSE), mean absolute error (MAE), model BIAS (BIAS), critical success index (CSI), and Pearson correlation (PC). These metrics provide a comprehensive aggregated point-by-point analysis of the CNN's performance with regards to the numerical error and the categorical accuracy. The mathematical equations for each are shown in Equation S3.

Similarly to Sperati et al. (2017), to verify the spatial consistency of the prediction generated by each of the methods, we also compare the pairwise correlation between all pairs of grid points for the predictions with the observations. When the pairwise correlation between a chosen model's grid points (e.g., the CNN) more closely matches the pairwise correlation for the ground truth grid points, it indicates a greater degree of correspondence in terms of spatial relationships in the ground truth.

3 Results

The U-Net CNN post-processed forecasts are compared against several methods. Figure 1 shows an example of a 96 h forecast of an extreme event that occurred on February 10, 2014 in the test set. The multilinear regression post-processing and West-WRF model overpredict over the highlighted heavy precipitation areas. Comparatively, the CNN qualitatively more closely resembles the observation patterns of the event as estimated by PRISM, especially within the heavy precipitation regions. For this case, it produces the lowest RMSE with respect to the PRISM ground-truth field, improving upon West-WRF by 33.9% and MOS by 8.1%. This is an example of CNN's ability to correct for spatial biases in the forecasts.

3.1 Discussion of Evaluation Metrics

The models are compared with respect to all the error metrics defined in Section 2.4: RMSE, MAE, PC, and CSI. All of the shown metrics and improvements were bootstrap sampled and produced with a 95% confidence interval to indicate if the results are statistically significant.

The CNN's overall RMSE aggregated over the 4 lead times (1-4 days) consistently outperformed climatology by 34.1-37.0%, West-WRF by 12.9-15.9%, and MOS by 7.4-8.9%. Similarly, the CNN outperformed both West-WRF and MOS for all 4 lead times with respect to Pearson correlation (PC) by 2.7-3.4% and 3.3-4.2%, respectively. Over the same period, the CNN improved upon West-WRF's CSI by 0.6-1.5%, with greater improvements ranging from 2.7% to 5.6% for lead times of 24 to 48 h. Note that we do not provide a complete set of improvement statistics for CLIM apart from RMSE since it is consistently 40-50% improved upon with regard to every metric.

Further, we analyze the performance of the models on the top 10% most heavy precipitation events. The CNN's overall RMSE/MAE over these events was reduced 19.8-21.0/17.7-18.3% and 8.8-9.7/5.4-6.2% compared to West-WRF and MOS respectively

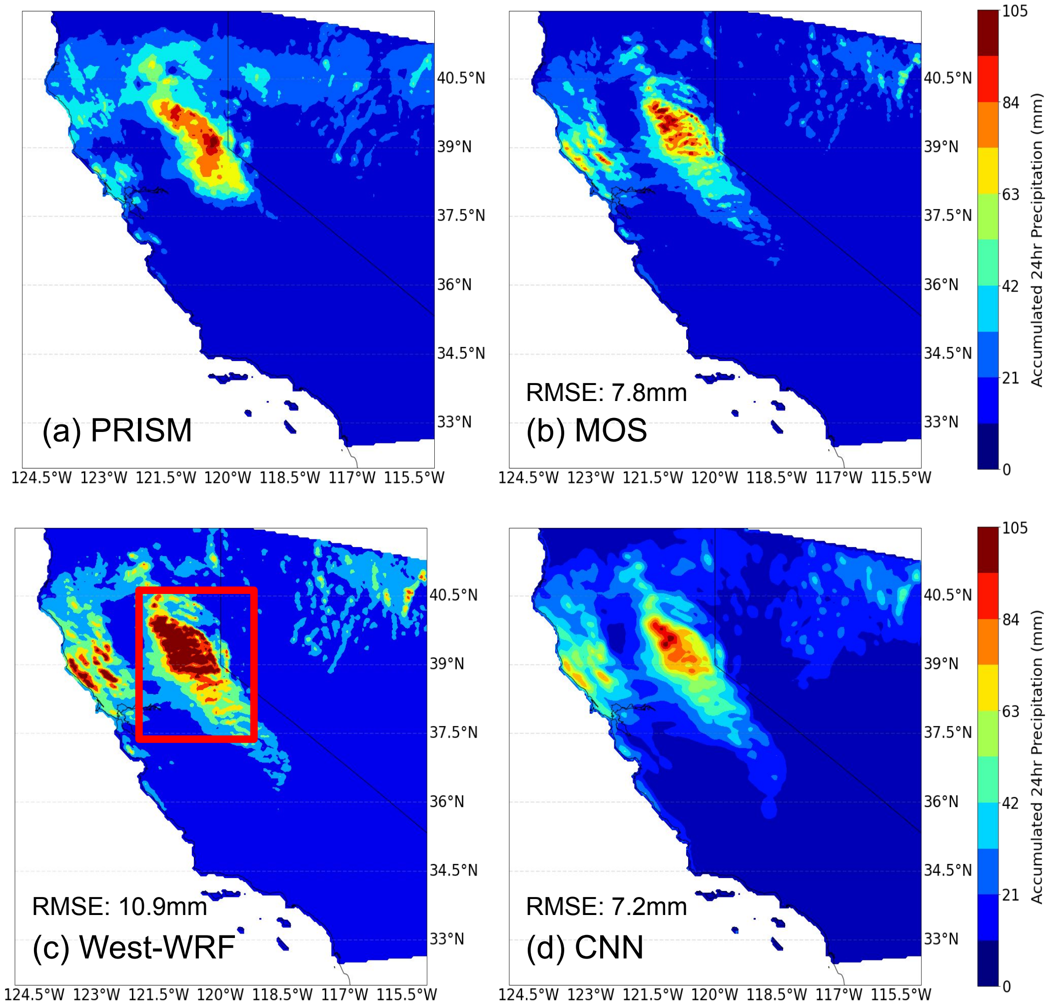


Figure 1. The 24-h accumulated precipitation on February 10, 2014 (test set) for (a) PRISM, (b) MOS, (c) West-WRF, and (d) CNN. The RMSE for each method with respect to the PRISM observation (a) is (b) 7.8 mm, (c) 10.9 mm, and (d) 7.2 mm. The highlighted region in red showcases an area of strong overprediction in West-WRF.

over all lead times of 1-4 days. Further, the CNN's PC over these events was improved by 4.9-5.5% and 4.2-4.7% compared to West-WRF and MOS.

Since the latter two metrics, PC and CSI, showcase the spatial and categorical accuracy of the methods, and RMSE summarizes the numerical accuracy, the CNN clearly outperformed the other post-processing and dynamical methods over all lead times with respect to spatial, categorical, and numerical accuracy aggregated over heavy precipitation and all events. These improvements are all shown to be statistically significant over a 95% confidence interval. The complete comparison for each error metric over each model is included in the supplemental information for all lead times.

These improvements are qualitatively consistent or better over similar dynamical baselines as cited in recent literature regarding machine learning-based post-processing methods. In Roulin and Vannitsem (2012), probabilistic techniques such as logistic regression are used to improve precipitation forecasts. Over the forecasting period, the MSE throughout the forecasting period is 5-15% better than the baseline dynamical method, which is consistent with the multilinear regression model presented in this study that is shown to be 28-31% inferior to the CNN in terms of MSE.

3.2 Temporal Evaluation of Models

We show some of the error metrics (RMSE, CRMSE, BIAS, PC) for each post-processing and dynamical method as a function of the lead time in Figure 2. This allows a more thorough examination of the propagation of error through increasing lead times.

Specifically, the RMSE is decomposed into bias, which reflect systematic errors, and CRMSE, which includes random errors and conditional biases, as indicated in Equation S1. Throughout the 4 lead times, the CNN consistently has the lowest CRMSE, as well as the highest Pearson correlation. In fact, the CNN is consistently able to add a day worth of predictive skill when compared to West-WRF (i.e. CNN error on day 4 is less than West-WRF error on day 3) in terms of RMSE, CRMSE, and PC. The BIAS fluctuates for each post-processing and forecasting method, but it is significantly lower than the CRMSE and contributes only marginally to the RMSE. This means that the CNN is able to improve the predictive ability of the dynamical model while minimally increasing the systematic errors (when compared to total RMSE).

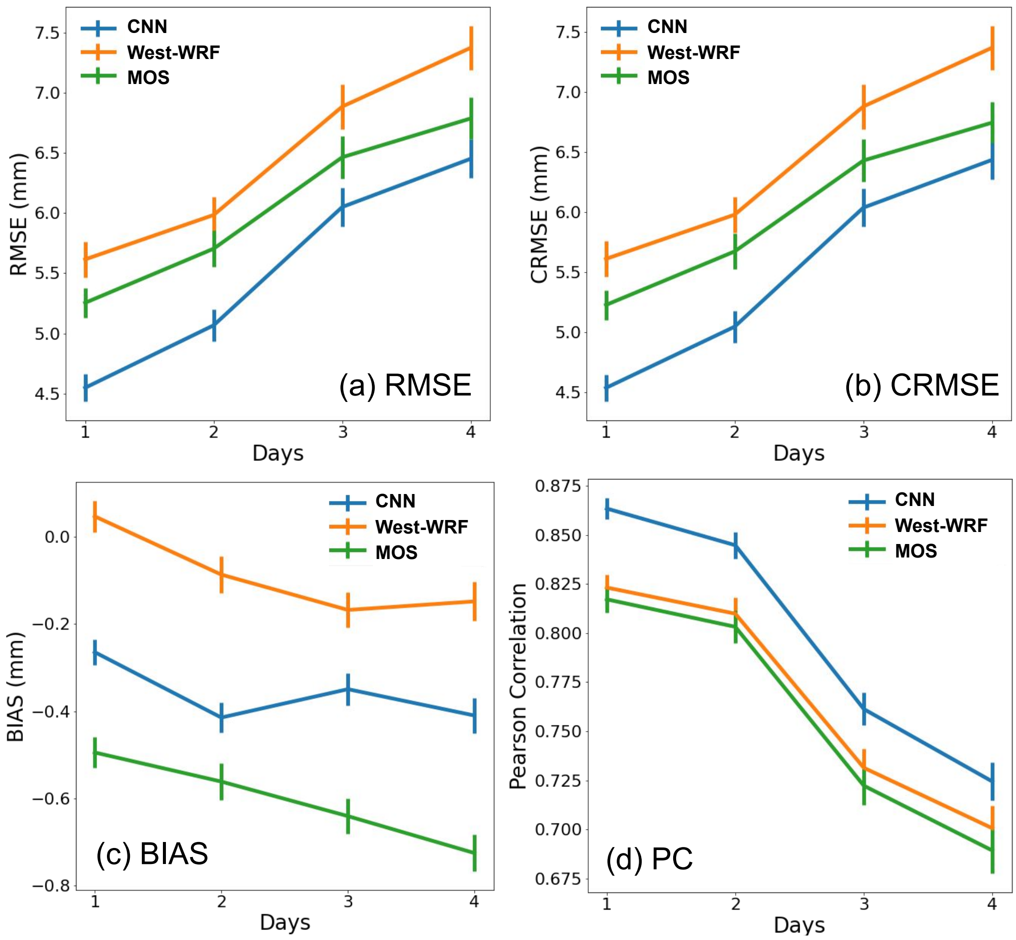


Figure 2. Pearson's correlation, CRMSE, RMSE, and BIAS for each model as a function of model lead time in days.

Further, we evaluate the rate of growth of the error metrics to evaluate the CNN's capabilities of producing longer-term forecasts and the scaling of the error as a function of lead time. A slower rate of growth of all error metrics would indicate a method that tracks better as a function of lead time. The average rate of growth for RMSE is significantly higher for West-WRF between days 2-4 as compared to CNN, with a reduction of 17.9% from day 3 to 4. Similarly, the average rate of decay for PC is reduced by 16.6% for the CNN as compared to West-WRF over days 2-4. Effectively, the CNN add more than one day of predictive skills to West-WRF, as for example is indicated by the CNN's RMSE at day 4, which is between the RMSE of West-WRF at days 2 and 3.

3.3 Spatial Evaluation of Models

The spatial patterns of improvement in error metrics as compared to West-WRF, aggregated over lead times of Day 1 to Day 4, is shown in Figure 3. The improvement in RMSE and MAE are consistently above 10%, with significant improvements of around 30-40% in the Sierra Nevada region. Similarly, the CNN improves upon West-WRF's Pearson correlation coefficient 5% or more, with significant improvements of 10-15% in southern California. The sharp decrease in correlation in the northern region and throughout the California Channel Islands is likely attributed to the CNN's documented weaknesses to domain boundaries due to spatial padding for convolution (Alsallakh et al., 2020). The CNN's improvements in CSI are largely mixed, with coastal California showing around 10% improvement over West-WRF. In southern California and Nevada, the West-WRF model outperforms the CNN by 15%. However, it is important to note that regions in which the CNN more significantly underperforms (the highlighted blue regions) account for only 9.2% of the total precipitation in the region (i.e., they are dry areas).

The spatial consistency of the generated precipitation field is also examined using a pairwise correlation plot (Sperati et al., 2017). This is an important aspect of the forecast evaluation because it explores the ability of the CNN to capture the spatial distribution of observed precipitation.

The pairwise correlation plot is shown in Figure 4 for both the CNN and West-WRF methods. With a perfect forecasting or post-processing method, we expect the correlation between each of the grid cells to match with the observation set, as shown by the 1:1 line in orange. The actual distribution of pairwise correlations between the CNN and West-WRF with respect to the PRISM is shown as a density plot. Qualitatively, it is noted that the CNN maintains the spatial attributes of the PRISM observations just as well as West-WRF by the fact that the spread is just as concentrated along 1:1 line. The higher coefficient of determination (R^2) of the CNN pairwise correlation plot indicates that the dispersion around the identity is lower than that of the West-WRF pairwise correlation plot. This indicates the CNN's superior spatial consistency with the PRISM ground truth as compared to West-WRF. Note that this analysis does not factor in the observational error.

4 Conclusions

The U-Net Convolutional Neural Network (CNN) architecture originally proposed by Ronneberger et al. (2015) and adapted in this study for precipitation prediction provides a computationally efficient and consistently accurate post-processing framework over different types of water years that outperforms competing machine learning and dynamical models. It provides superior spatial consistency and numerical accuracy over all lead times as summarized by the 12.9-15.9% improvement in root-mean-square error (RMSE) over the Western Weather Research and Forecasting model and 7.4-8.9% improvement over Model Output Statistics. It also displays a reduce rate of error growth such as RMSE and Pearson's correlation as lead times increase, which effectively results in more than a day of additional predictive skill with respect to a dynamical model. Ad-

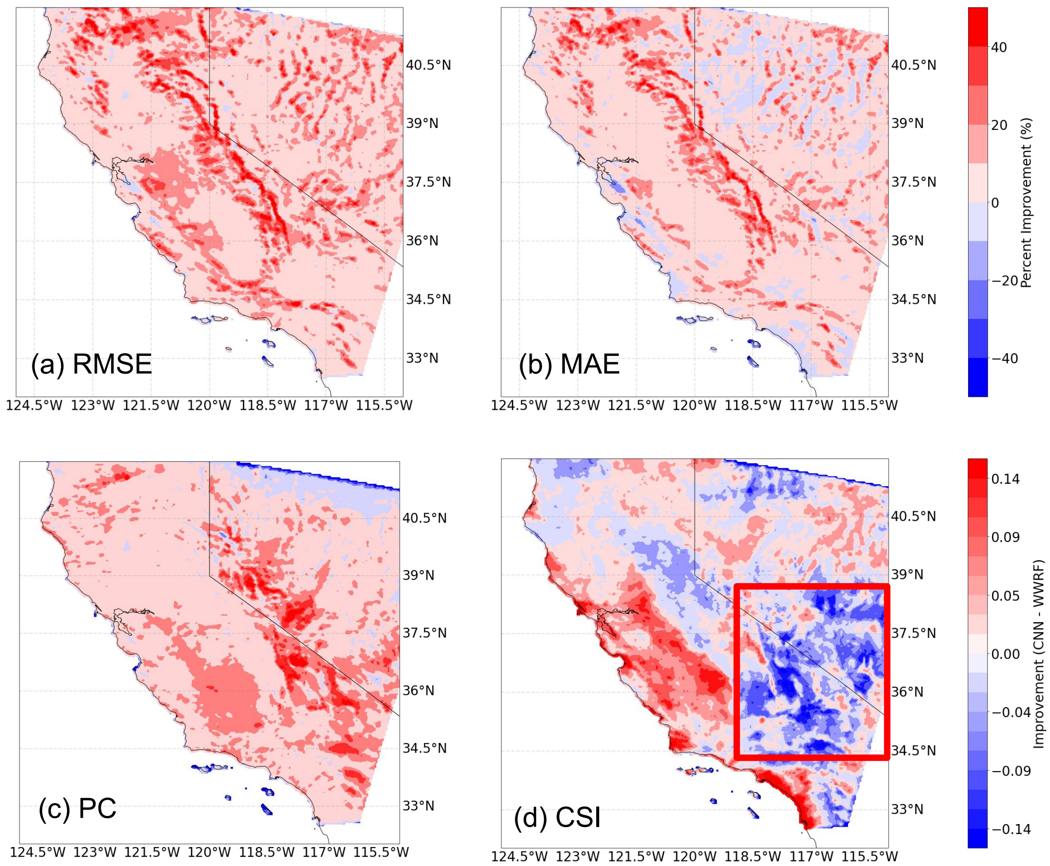


Figure 3. The CNN's improvement/degradation in RMSE, MAE, PC and CSI as compared to the West-WRF regional model aggregated over all lead times (1-4 days). Highlighted region shows an area of severe reduction in CSI for the CNN, see discussion in text.

ditionally, the CNN outperforms the other methods for the prediction of the top 10% precipitation events. This demonstrates a consistent and reliable post-processing framework that improves upon spatial and temporal biases over dynamical models and other post-processing methods over the western US. Future work includes examining the temporal association between day-to-day forecasts using recurrent neural networks or transformers along with an encoding convolutional neural network. The Convolutional Long Short-Term Memory layer developed by Shi et al. (2015) provides a promising avenue to explore this further. Additional methods to rectify the class imbalance can be explored, such as data augmentation.

Acknowledgments

This research was supported by USACE FIRO grant W912HZ1520019 and CDWR AR Program grant 4600013361.

References

Alessandrini, S., Delle Monache, L., Sperati, S., & Nissen, J. (2015). A novel application of an analog ensemble for short-term wind power forecasting. *Renewable Energy*, *76*, 768-781. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0960148114007915` doi: https://doi.org/10.1016/

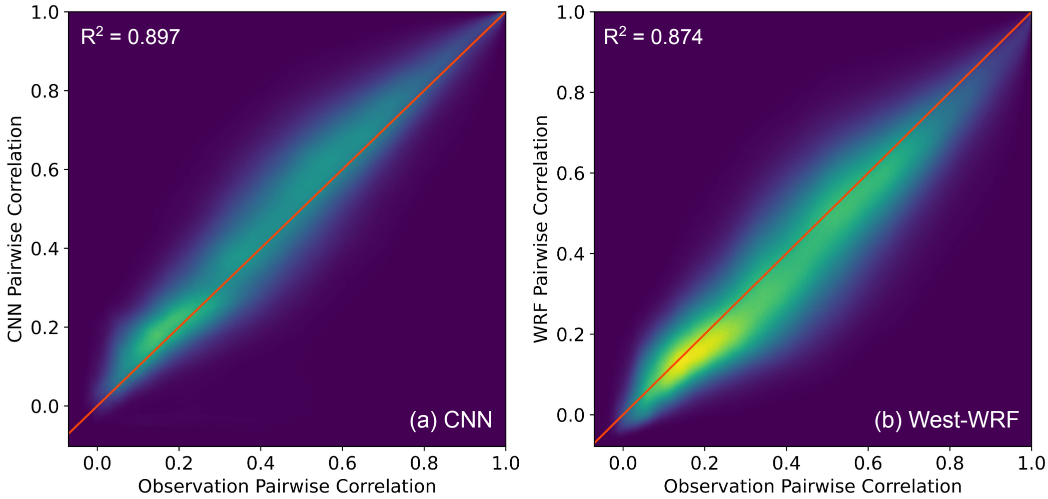


Figure 4. The CNN's (left) and West-WRF (right) pairwise correlation plot with the PRISM observations. The coefficient of determination (R^2) is shown in the upper left of both panels. The orange line denotes a perfect correspondence in observation and model pairwise correlation between grid points.

j.renene.2014.11.061

Alsallakh, B., Kokhlikyan, N., Miglani, V., Yuan, J., & Reblitz-Richardson, O. (2020). Mind the pad–cnns can develop blind spots. *arXiv preprint arXiv:2010.02178*.

AMS. (2019). *Glossary of meteorology.* (Retrieved from `http://glossary.ametsoc.org/wiki/atmospheric_river` at Jun 9, 2021 2:39 PM)

Anelli, V. W., Di Noia, T., Di Sciascio, E., Pomo, C., & Ragone, A. (2019). On the discriminative power of hyper-parameters in cross-validation and how to choose them. In *Proceedings of the 13th acm conference on recommender systems* (pp. 447–451).

Cervone, G., Clemente-Harding, L., Alessandrini, S., & Delle Monache, L. (2017). Short-term photovoltaic power forecasting using artificial neural networks and an analog ensemble. *Renewable Energy*, *108*, 274-286. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0960148117301386` doi: https://doi.org/10.1016/j.renene.2017.02.052

Chapman, W., Subramanian, A., Delle Monache, L., Xie, S., & Ralph, F. (2019). Improving atmospheric river forecasts with machine learning. *Geophysical Research Letters*, *46*(17-18), 10627–10635.

Chapman, W. E., Subramanian, A. C., Xie, S.-P., Sierks, M. D., Ralph, F. M., & Kamae, Y. (2021). Monthly modulations of enso teleconnections: Implications for potential predictability in north america. *Journal of Climate*, 1–71.

Chollet, F., et al. (2015). *Keras.* `https://github.com/fchollet/keras`. GitHub.

Collins, M., & Allen, M. R. (2002). Assessing the relative roles of initial and boundary conditions in interannual to decadal climate predictability. *Journal of Climate*, *15*(21), 3104–3109.

Corringham, T. W., Ralph, F. M., Gershunov, A., Cayan, D. R., & Talbot, C. A. (2019). Atmospheric rivers drive flood damages in the western united states. *Science advances*, *5*(12), eaax4631.

Daly, C., Halbleib, M., Smith, J. I., Gibson, W. P., Doggett, M. K., Taylor, G. H., . . . Pasteris, P. P. (2008). Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous united states.

International Journal of Climatology: a Journal of the Royal Meteorological Society, *28*(15), 2031–2064.

Daly, C., Slater, M. E., Roberti, J. A., Laseter, S. H., & Swift Jr, L. W. (2017). High-resolution precipitation mapping in a mountainous watershed: ground truth for evaluating uncertainty in a national precipitation dataset. *International Journal of Climatology*, *37*, 124–137.

Delle Monache, L., Eckel, F. A., Rife, D. L., Nagarajan, B., & Searight, K. (2013). Probabilistic weather prediction with an analog ensemble. *Monthly Weather Review*, *141*(10), 3498–3516.

Gibson, P. B., Waliser, D. E., Lee, H., Tian, B., & Massoud, E. (2019). Climate model evaluation in the presence of observational uncertainty: Precipitation indices over the contiguous united states. *Journal of Hydrometeorology*, *20*(7), 1339–1357.

Glahn, H. R., & Lowry, D. A. (1972). The use of model output statistics (mos) in objective weather forecasting. *Journal of Applied Meteorology and Climatology*, *11*(8), 1203–1211.

Hayatbini, N., Kong, B., Hsu, K.-l., Nguyen, P., Sorooshian, S., Stephens, G., ... Ganguly, S. (2019). Conditional generative adversarial networks (cgans) for near real-time precipitation estimation from multispectral goes-16 satellite imageries—persiann-cgan. *Remote Sensing*, *11*(19), 2193.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*, 1097–1105.

Kumar, A., & Hoerling, M. P. (1998). Annual cycle of pacific–north american seasonal predictability associated with different phases of enso. *Journal of Climate*, *11*(12), 3295–3308.

Louka, P., Galanis, G., Siebert, N., Kariniotakis, G., Katsafados, P., Pytharoulis, I., & Kallos, G. (2008). Improvements in wind speed forecasts for wind power prediction purposes using kalman filtering. *Journal of Wind Engineering and Industrial Aerodynamics*, *96*(12), 2348–2362.

Martin, A. C., Ralph, F. M., Wilson, A., DeHaan, L., & Kawzenuk, B. (2019). Rapid cyclogenesis from a mesoscale frontal wave on an atmospheric river: Impacts on forecast skill and predictability during atmospheric river landfall. *Journal of Hydrometeorology*, *20*(9), 1779–1794.

Nicolis, C., & Nicolis, S. C. (2007). Return time statistics of extreme events in deterministic dynamical systems. *EPL (Europhysics Letters)*, *80*(4), 40003.

ODonnell, A., Hubbard, T., Nadeau, L., Delaney, C., Hartman, R., Mendoza, J., ... Corringham, T. (2020). Estimating benefits of forecast-informed reservoir operations (firo): Lake mendocino case-study and transferable decision support tool. In *Agu fall meeting 2020*.

Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.

PRISM Climate Group, O. S. U. (2004). *Prism data.* (Retrieved from `http://prism.oregonstate.edu`)

Ralph, F. M., Rutz, J. J., Cordeira, J. M., Dettinger, M., Anderson, M., Reynolds, D., ... Smallcomb, C. (2019). A scale to characterize the strength and impacts of atmospheric rivers. *Bulletin of the American Meteorological Society*, *100*(2), 269–289.

Rasp, S., & Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, *146*(11), 3885–3900.

Richardson, L. F. (1922). *Weather prediction by numerical process.* Cambridge university press.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241).

Roulin, E., & Vannitsem, S. (2012). Postprocessing of ensemble precipitation predictions with extended logistic regression based on hindcasts. *Monthly weather review*, *140*(3), 874–888.
Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., & Woo, W.-c. (2015). Convolutional lstm network: A machine learning approach for precipitation nowcasting. *arXiv preprint arXiv:1506.04214*.
Sperati, S., Alessandrini, S., & Delle Monache, L. (2017). Gridded probabilistic weather forecasts with an analog ensemble. *Quarterly Journal of the Royal Meteorological Society*, *143*(708), 2874–2885.
Steinhoff, D., Kawzenuk, B., Weihs, R., Reynolds, D., DeHaan, L., Martin, A., & Delle Monache, L. (2020). Nrt wy2020 post-season report.
Tao, Y., Gao, X., Hsu, K., Sorooshian, S., & Ihler, A. (2016). A deep neural network modeling framework to reduce bias in satellite precipitation products. *Journal of Hydrometeorology*, *17*(3), 931 - 945. Retrieved from `https://journals.ametsoc.org/view/journals/hydr/17/3/jhm-d-15-0075_1.xml` doi: 10.1175/JHM-D-15-0075.1
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.
Vannitsem, S., & Ghil, M. (2017). Evidence of coupling in ocean-atmosphere dynamics over the north atlantic. *Geophysical Research Letters*, *44*(4), 2016–2026.

fig1_final.pdf.

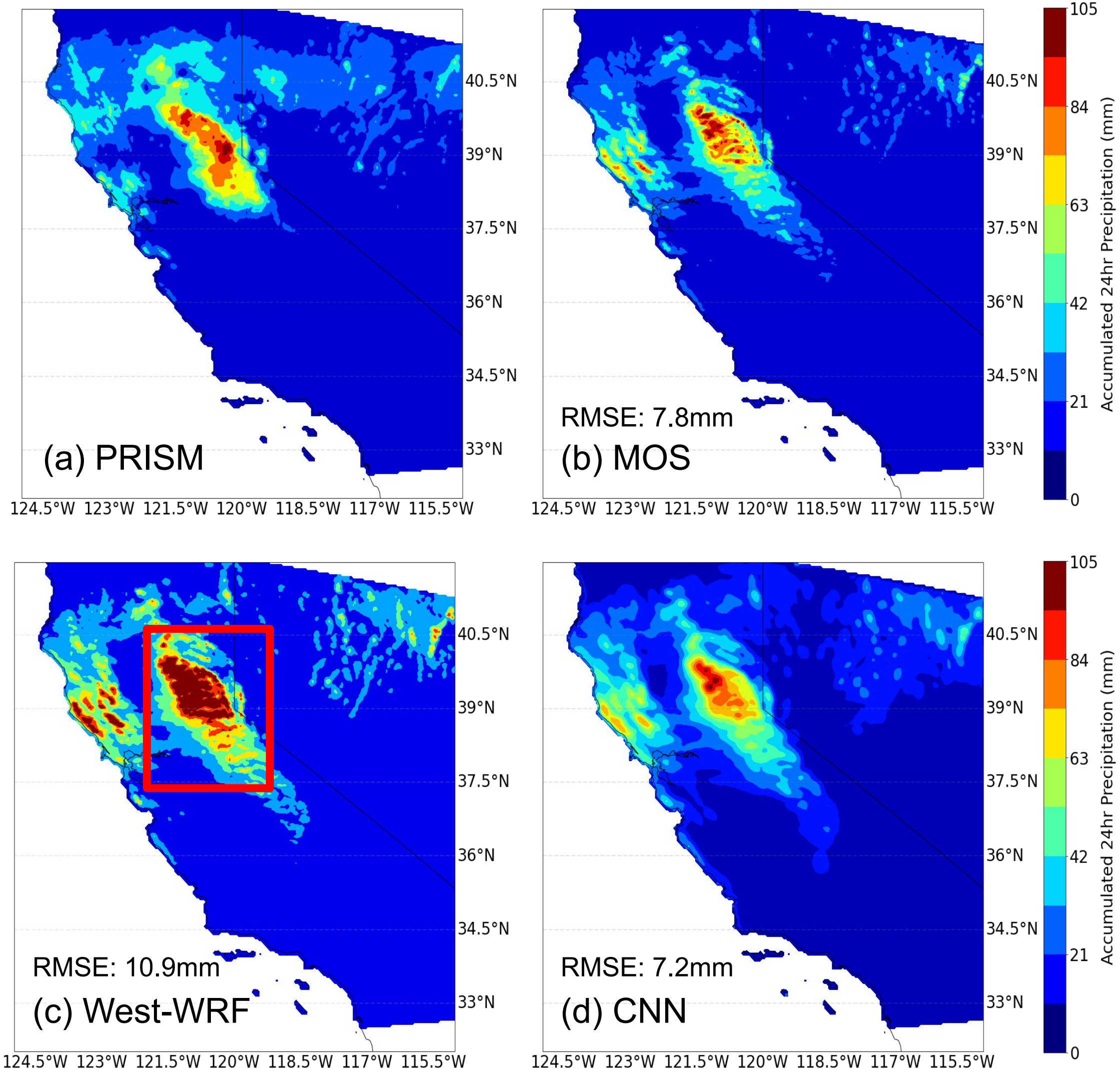
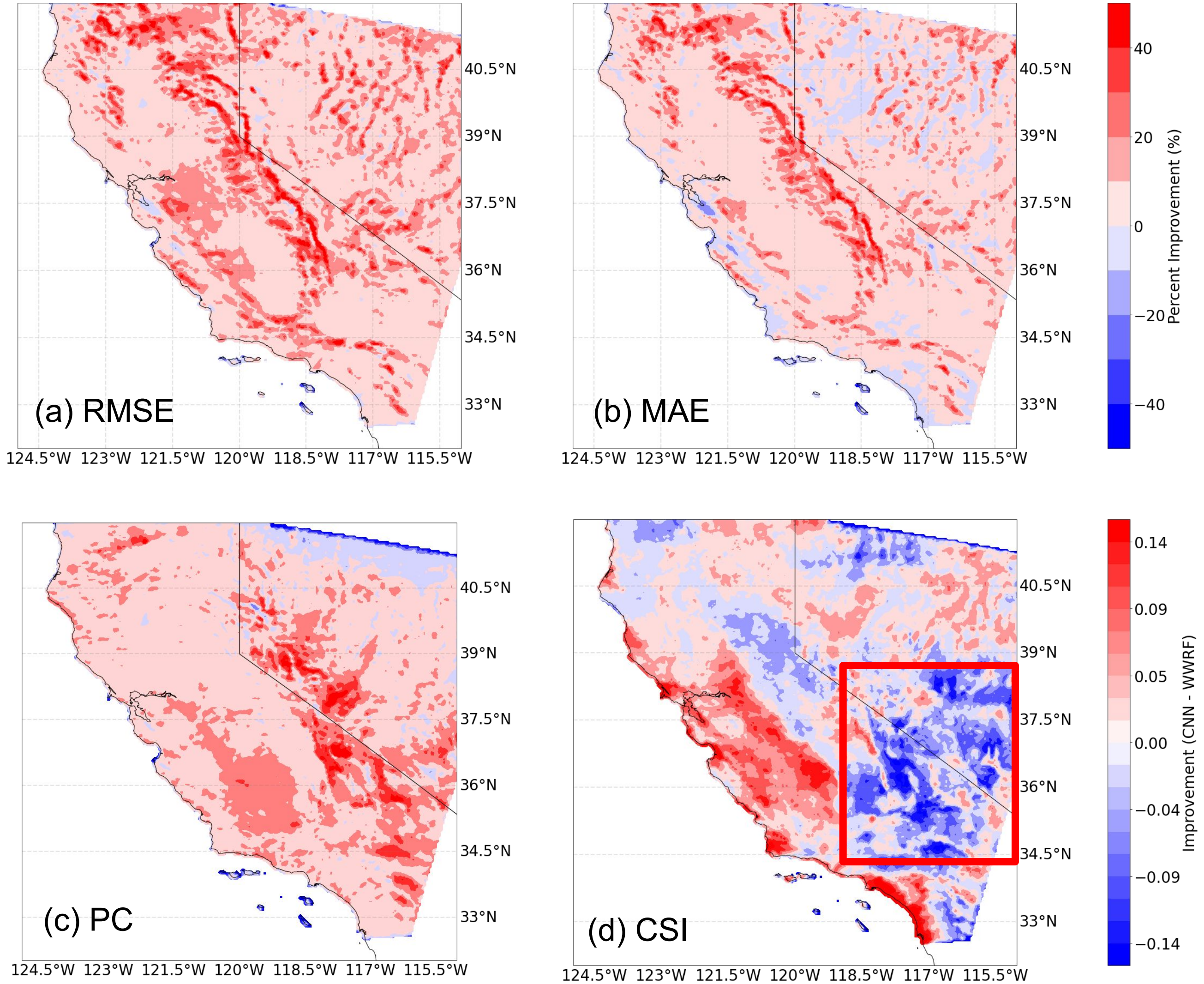
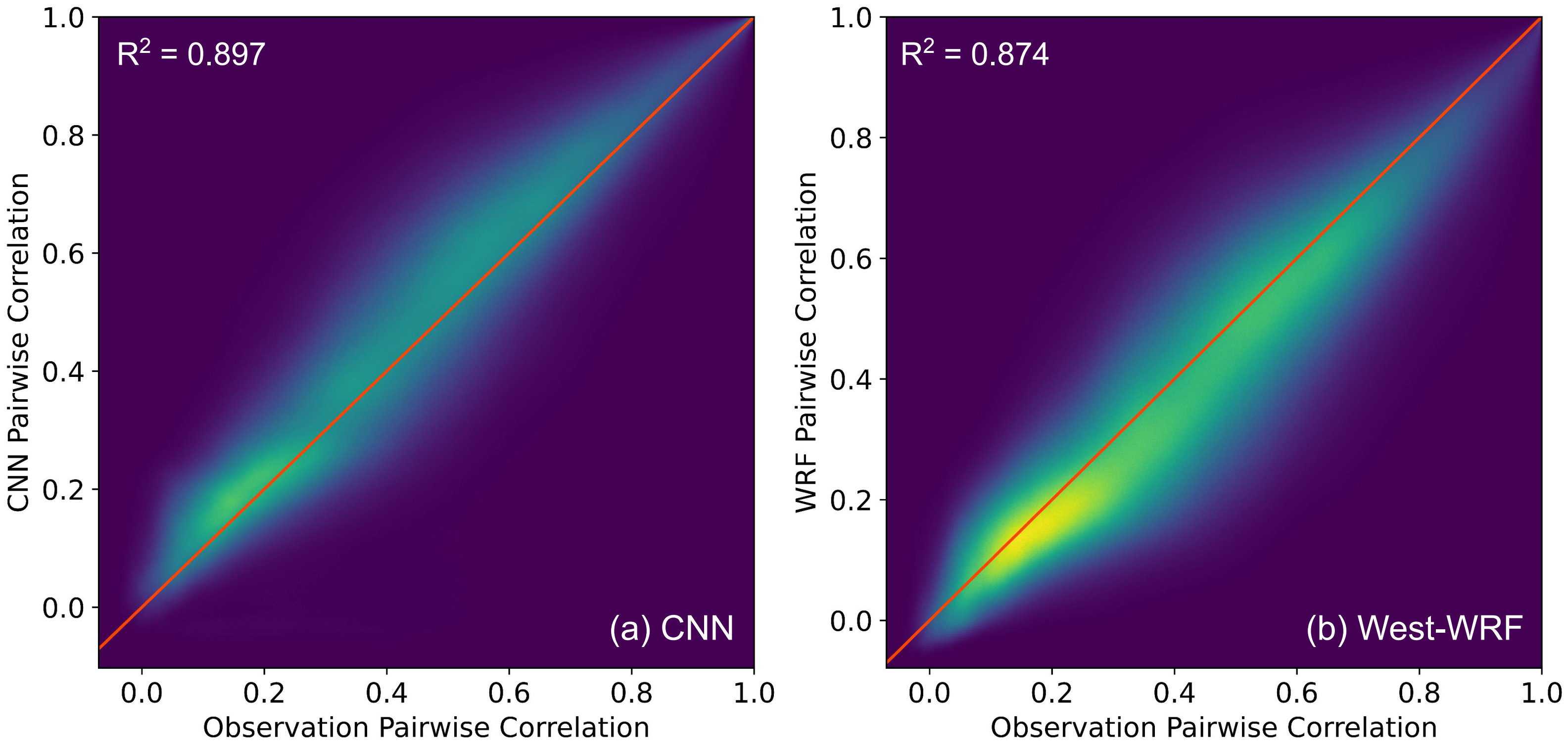

(a) PRISM
RMSE: 7.8mm
(b) MOS
RMSE: 10.9mm
(c) West-WRF
RMSE: 7.2mm
(d) CNN
Accumulated 24hr Precipitation (mm)
0
21
42
63
84
105
124.5°W 123°W 121.5°W 120°W 118.5°W 117°W 115.5°W
33°N
34.5°N
36°N
37.5°N
39°N
40.5°N

fig3_pdf.pdf.



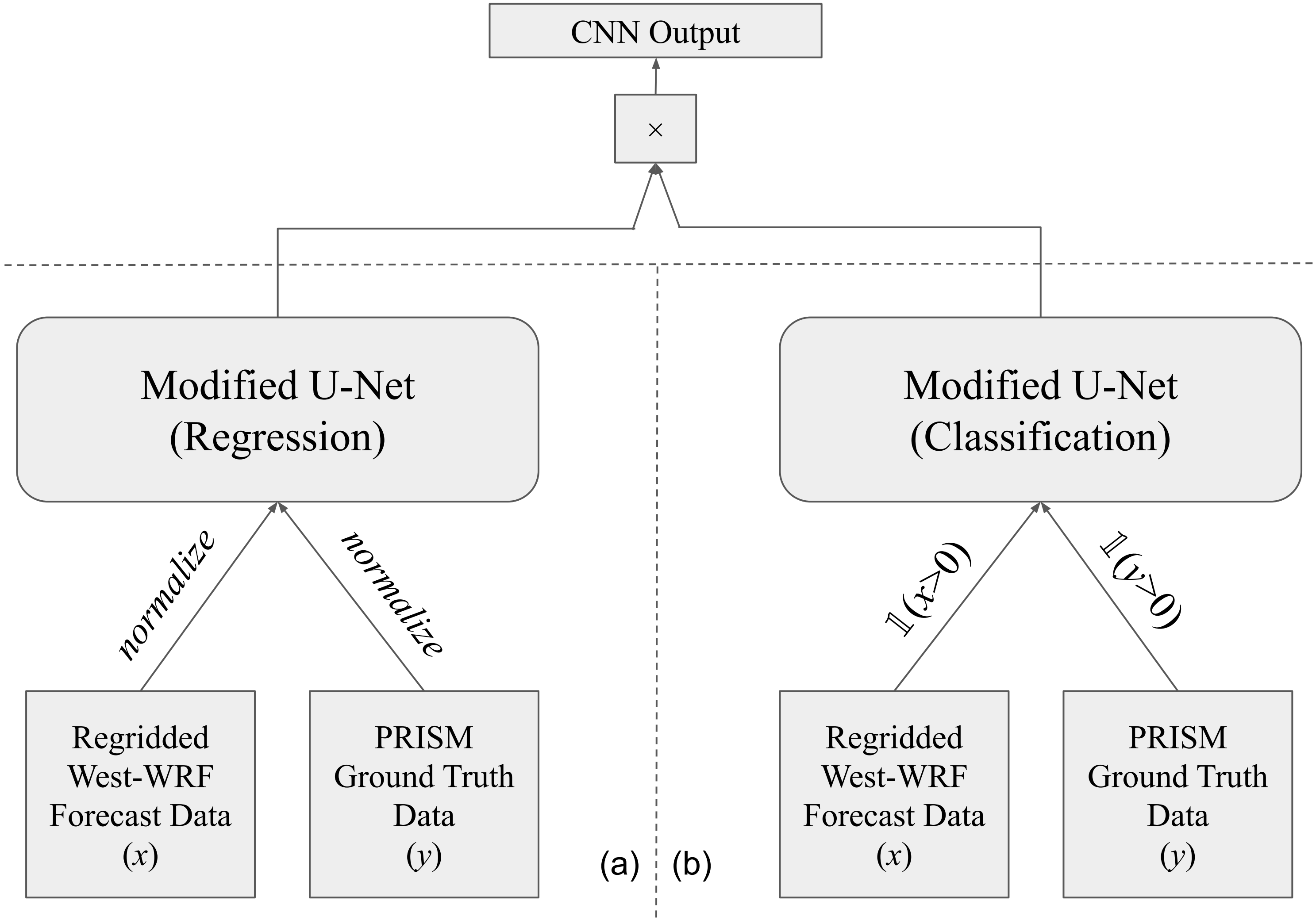
(a) RMSE
(b) MAE
(c) PC
(d) CSI
124.5°W 123°W 121.5°W 120°W 118.5°W 117°W 115.5°W
40.5°N
39°N
37.5°N
36°N
34.5°N
33°N
Percent Improvement (%)
40
20
0
−20
−40
Improvement (CNN - WWRF)
0.14
0.09
0.05
0.00
−0.04
−0.09
−0.14

fig4_pdf.pdf.



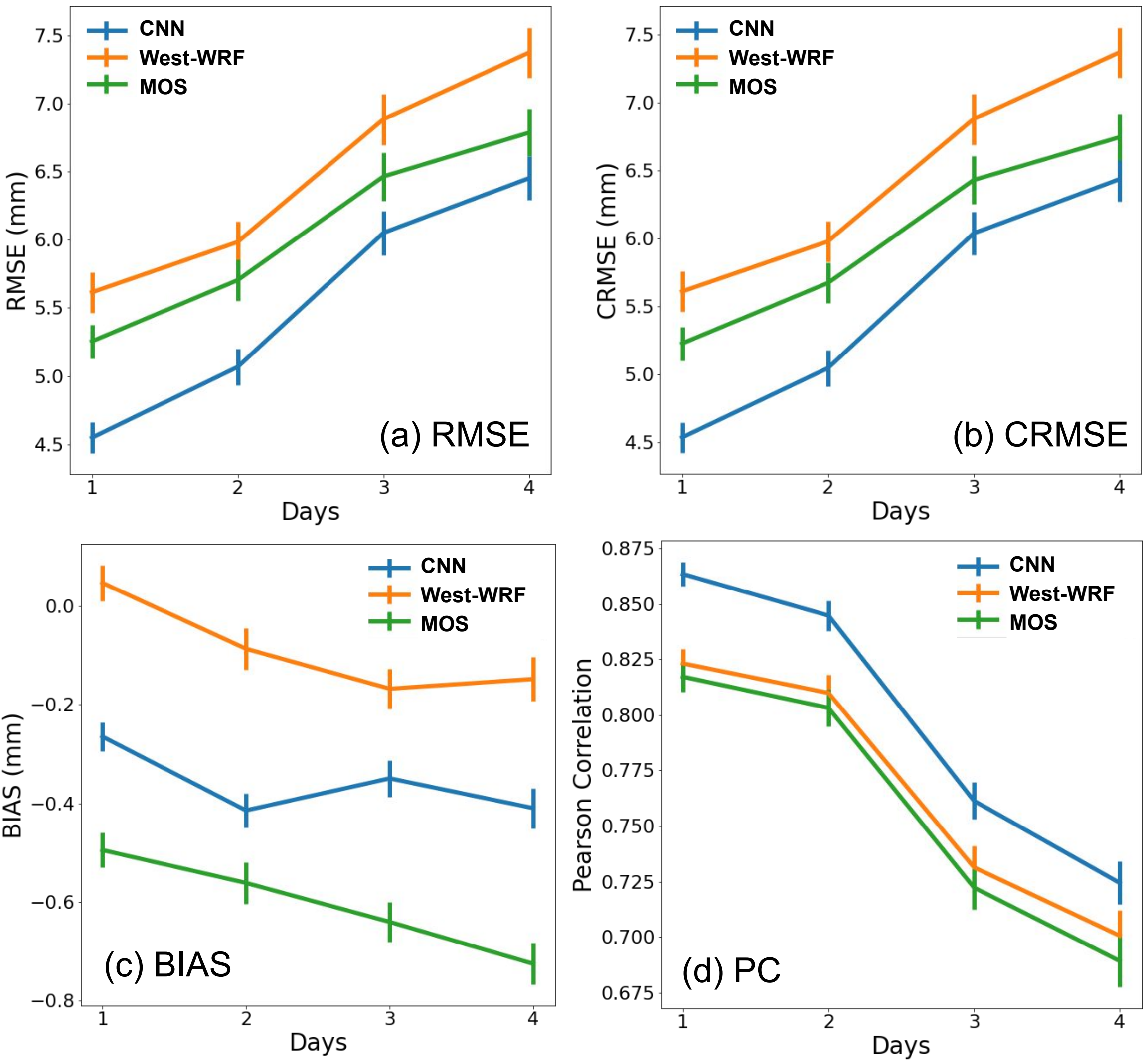
R² = 0.897
CNN Pairwise Correlation
Observation Pairwise Correlation
(a) CNN
R² = 0.874
WRF Pairwise Correlation
Observation Pairwise Correlation
(b) West-WRF
1.0
0.8
0.6
0.4
0.2
0.0

fig5.pdf.



CNN Output
×
Modified U-Net
(Regression)
Modified U-Net
(Classification)
normalize
normalize
$\mathbb{1}(x>0)$
$\mathbb{1}(y>0)$
Regridded
West-WRF
Forecast Data
(x)
PRISM
Ground Truth
Data
(y)
Regridded
West-WRF
Forecast Data
(x)
PRISM
Ground Truth
Data
(y)
(a)
(b)

fig2_pdf.pdf.



CNN
West-WRF
MOS
(a) RMSE
RMSE (mm)
Days
(b) CRMSE
CRMSE (mm)
(c) BIAS
BIAS (mm)
(d) PC
Pearson Correlation