Predicting Winter California Precipitation with Convolutional Neural Networks

Anthony Chiado[1], Kristian Olsson[1], Luke Rohlwing[1], Michael Vaden[1], Antonios Mamalakis[1, 2]

[1] School of Data Science, University of Virginia, Charlottesville, VA

[2] Department of Environmental Sciences, University of Virginia, Charlottesville, VA (e-mail: npa4tg@virginia.edu)

***Abstract* — Predicting winter precipitation in California is crucial for policy decisions, ecosystem health, and inhabitants' well-being. However, high variability and difficulty in prediction pose significant challenges. This study explores the potential of using a Convolutional Neural Network (CNN) trained on global summer sea surface temperatures (July-October) to forecast winter precipitation (November-March) across northern, central, and southern California. We leverage data from the Community Earth System Model 2 (CESM2) climate simulations for pre-training the CNNs and observational records to fine-tune the CNNs for real-world predictions. Testing on historical data from 1990-2021, the CNNs achieved R^2 values of 0.089, 0.311, and 0.336 for the northern, central, and southern regions, respectively. The CNNs outperformed a baseline linear regression model, which had R^2 values of 0.005, 0.029, and 0.108. We also generated saliency maps to identify important sources of predictability around the globe. Our study provides evidence that predictions of California's hydroclimate can be enhanced through the combination of deep learning and data from large-ensemble climate simulations.**

I. INTRODUCTION

A. Motivation

California experiences both prolonged droughts and extreme winter precipitation, significantly impacting policy decisions, ecosystems, and inhabitants. Predicting these events is crucial for effective water management, disaster preparedness, and agricultural planning. However, the high variability and inherent difficulty of forecasting California's winter precipitation pose a significant challenge.

Existing approaches, while providing valuable insights, often exhibit limited predictive performance, particularly in northern and central California. This research seeks to address this challenge by developing a novel and potentially more accurate method for predicting winter precipitation levels across California.

B. Summary of Existing Literature

Predicting California's winter precipitation (November-March) remains a challenge, despite ongoing research efforts. While previous methods have achieved an R^2 of around 0.4 for southwestern US [5], there is a need for improvement in predictive accuracy, particularly for northern and central California. Existing approaches utilize various climate indices and statistical models for prediction. These include:

- El Niño-Southern Oscillation (ENSO): Studies like Allen et al. (2017) [1] explore how El Niño can influence California precipitation in a warming climate. However, ENSO's influence can be complex and may not fully explain regional variations.
- Sea Surface Temperatures (SSTs): Research by Liu et al. (2018) [3] investigates the potential of autumnal sea surface salinity patterns to predict winter precipitation in the southwestern U.S. Similarly, Mamalakis et al. (2018) [4] explore the link between interhemispheric teleconnections and winter precipitation predictability.
- Statistical Modeling: Stevens et al. (2020) [5] utilize graph-guided regularized regression techniques to analyze Pacific Ocean climate variables for improved prediction of winter precipitation in the southwestern U.S.

While these methods offer valuable insights, they may not fully capture the intricate relationships between global climate patterns and regional precipitation. Additionally, traditional statistical models may struggle to capture complex non-linear patterns within the data. This is where

our study proposes a novel approach using a CNN to overcome these limitations.

C. *Novelty of Current Work*

This study introduces a novel approach utilizing a CNN trained on global summer sea surface temperatures (July-October) to predict winter precipitation (November-March) across northern, central, and southern California. This method leverages the ability of CNNs to identify intricate patterns in large datasets, potentially leading to more accurate predictions compared to traditional statistical models.

Furthermore, by using 100 climate simulations (each simulation consists of 74 years) from the Community Earth System Model 2 (CESM2), our research offers a robust training foundation for CNNs. Next, we use fine-tuning of the model to predict in the real-world, an approach that holds the potential for significant advancements in the task of predicting California's winter precipitation.

This research not only aims to improve predictive accuracy but also establishes CNNs and more generally, neural networks, as a promising method for future research in California's winter precipitation forecasting. Building upon this approach, researchers can explore other neural networks and architectures to further enhance predictive capabilities.

II. METHODS

Our methodology involves processing and preparing two key datasets before building the CNN: sea surface temperatures and precipitation data. Both datasets encompass monthly observations of the features across 100 climate simulations from the Community Earth System Model 2 (CESM2) spanning 1940-2013.

A. *Sea Surface Temperatures (SST) Data*

In preparation for model training, we obtain SST data from 100 CESM2 simulations, represented in Kelvin degrees. We compute the average values for the summer months (July to October) for each latitudinal and longitudinal grid point and for each simulation. The SST series are then detrended for each latitudinal and longitudinal point and for each simulation, to remove any linear trends. By consolidating SST data this way, we create a comprehensive dataset conducive to training our predictive model for winter precipitation levels in California.

A subset of this data is used to represent the Nino3.4 index, calculated by averaging the SSTs within this specific Pacific region (see red box in Figure 1), weighting the area by latitude with a cosine function. This data is used to create a baseline linear model that we improve upon using a CNN.

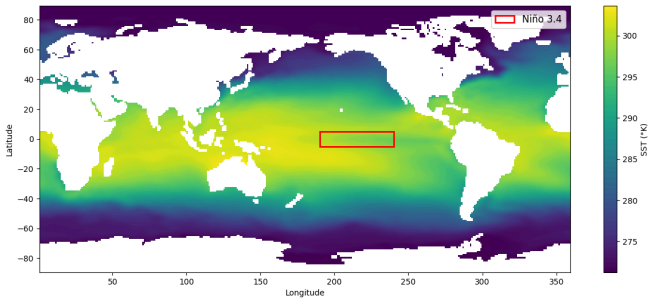

Figure 1. SST (K) Data for 1950 with Nino 3.4 index

Lastly, the real-world SST data is extracted from the COBE-SST 2 model developed by the Japanese Meteorological Center (JMA). This data is prepared similarly to the CESM2 SST data with the additional step of interpolating the coordinates to match the grid of the CESM2 before detrending. SST observations are limited to the years 1950-2021, chosen for its reliability during this period.

B. *Precipitation Data*

To prepare our response variable precipitation, we begin by obtaining precipitation data from all CESM2 simulations represented in millimeters per day. The precipitation data are then averaged over only the winter months of November through March within each simulation. This aggregation process yields a singular data point of total winter precipitation for each year within every simulation, labeled as PRECT. Subsequently, we calculate the weighted average of the data over each of the three defined regions of California: northern, central, and southern (Figure 2). We consider three California subregions to account for differences in the geography of California, and for variations in precipitation patterns across the state. Finally, to ensure data integrity, we detrend the aggregated precipitation values across the selected regions for each simulation (Figure 3),

removing any linear trends that may affect predictive accuracy.

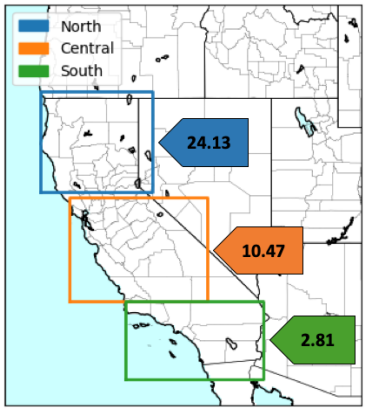

Figure 2. Average PRECT (mm/day) for each California region in 1950

For real-world precipitation data, we utilize PRECT data from the NCEP/Climate Prediction Center. We average these values for the winter months of November through March. Next, we calculate the weighted average of total winter precipitation for each year across the three California regions. The real-world precipitation dataset spans from 1950 to 2021, a period chosen for its increased reliability in observational data similar to real-world SST data.

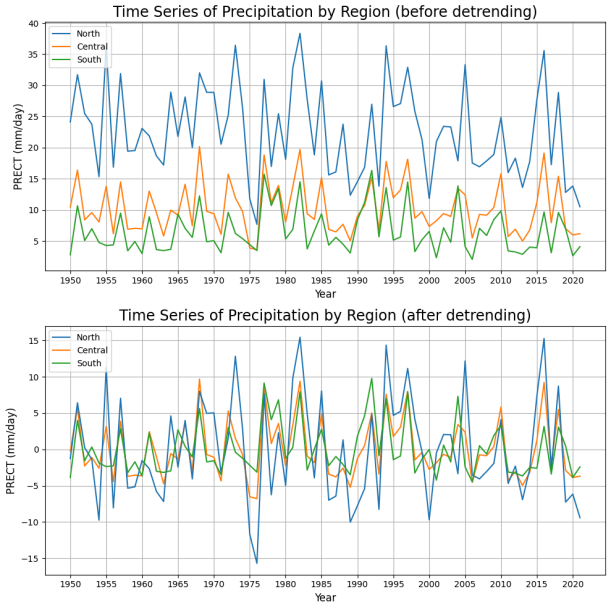

Figure 3. Time Series of Real-World Precipitation (mm/day) by Region, before and after detrending

C. Convolutional Neural Network (CNN) Model

The proposed CNN architecture introduces a sophisticated framework for addressing the challenges inherent in predicting precipitation levels across California. Leveraging convolutional, pooling, and fully connected layers, the CNN undergoes iterative training loops to optimize its parameters, minimizing prediction errors and enhancing its predictive capabilities. Dropout regularization techniques are incorporated to mitigate overfitting risks, ensuring the model's robustness to complex patterns in precipitation data. Additionally, hyperparameter tuning of the learning rate, dropout probabilities, and activation functions further refine the model's performance.

More specifically, the proposed CNN architecture includes a series of convolutional, pooling, and fully connected layers. The model consists of three convolutional layers, each followed by max-pooling operations to downsample the feature maps and extract essential features. The first convolutional layer employs 32 filters with a kernel size of 3x3 and a stride of 3, followed by rectified linear unit (ReLU) activation and 2D dropout regularization to prevent overfitting. Subsequently, the second convolutional layer reduces the number of filters to 16 while maintaining similar kernel size and stride parameters, further enhancing feature extraction. A third convolutional layer with 8 filters refines the learned features before transitioning to fully connected layers for prediction. There are three fully connected layers. The first two are followed by a ReLU activation. Each of these layers has decreasing dimensions, leading to a single output node, representing the predicted precipitation level. Dropout regularization is applied to the fully connected layers to prevent co-adaptation of neurons and improve model generalization. This architecture results in a model with 12,796 parameters.

Using the architecture in Figure 4, we deploy three distinct models tailored to predict precipitation levels within the specific regions of California. These models are individually trained with different hyperparameters optimized for their respective regions: northern, central, and southern California. This approach allows us to account for the unique climatic nuances and precipitation patterns characteristic of each region, optimizing the predictive accuracy of our models.

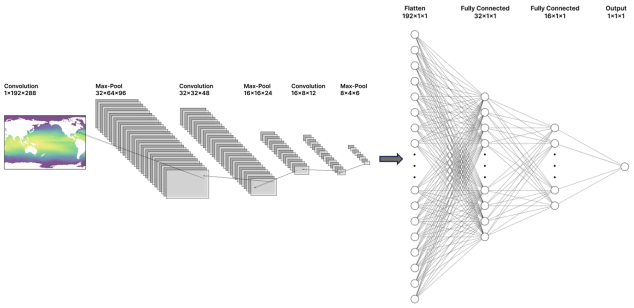

Figure 4. CNN Architecture Schematic

For our model training and evaluation, the dataset is split into training, validation, and test sets, with 80%, 10%, and 10% of the total data, respectively. During training, we employ early stopping mechanisms with a Mean Squared Error (MSE) loss function to prevent overfitting and optimize the model performance. Subsequently, the trained model is then tested on real-world data to test its predictive power outside of the simulated environment. Before testing on real-world data, we fine-tuned the final fully-connected layer of the CNN using observations from the years 1950-1989. We then tested on the years 1990-2021. The results can be seen in Table 1. This comprehensive approach to training and testing ensures the reliability and generalizability of our predictive model for precipitation forecasting.

Furthermore, to gain insights into the decision-making process of our models, we employ saliency maps. Saliency maps provide visualizations that highlight the most influential regions within the input data that contribute to the model's predictions. By analyzing these maps, we can discern which oceans basins or patterns the model prioritizes when making predictions for each California region.

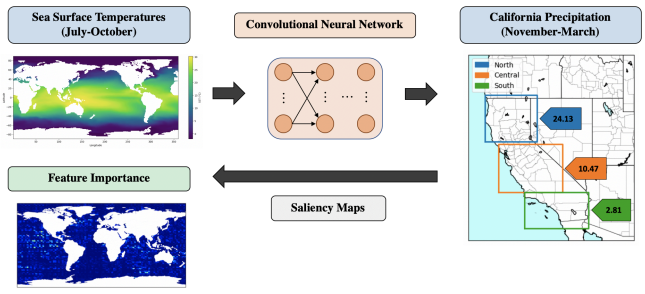

Figure 5. CNN Modeling Process with Saliency Maps

III. RESULTS

In general, our findings indicate that CNNs can be successfully applied to the problem at hand. On simulated data, CNNs outperformed the baseline linear regression model in all California regions (see Table 1). Likewise, when trained and tested on real-world data, the CNNs significantly outperformed the linear regression model. The CNN for the central region explained 31.1% of the variance in precipitation, compared to only 2.9% for the linear model. Similarly, the CNN models outperformed their linear regression counterparts by a large margin for the other regions (see Table 1).

Furthermore, we highlight that the use of fine-tuning on real-world data greatly improved predictive performance compared to if one was to use the pretrained CNN and predict in the real world without fine-tuning. This indicates that the simulated data can capture some of the patterns that occur within the real system, but certain aspects may not be fully represented. However, by fine-tuning the model on historical data, we were able to train the model to perform well outside of the simulated data (see Figure 6). Across all models and all datasets, we found that the northern region is the most difficult to predict and that the southern region is the easiest to predict. This aligns with previous research on the California hydroclimate [1].

The northern California model captured 0.089 of the variability, compared to the linear El Niño model which captures 0.005 (see Table 1). Looking at our specific predictions (see Figure 6), we captured the higher levels of precipitation well in 1997. The saliency map (see Figure 7) shows that our prediction for 1997 precipitation was heavily influenced by the sea surface temperatures in the Indian Ocean and the Pacific Ocean. The saliency map shows the importance of the El Niño, but also captured more complex patterns and features across the world's oceans.

The central California model captured 0.311 of the variance in the real-world test data, compared to the linear model that captured only 0.029 of the variance. Looking at the real-world precipitation in the central region, the CNN predicted 2015 with a low error (see Figure 6). The saliency map for this prediction (Figure 8) shows that the CNN captured some of the effects of El Niño, as well as data from the northwestern Indian Ocean, and the Southern Ocean. This indicates that the model learned more complex

non-linear relationships beyond a simple linear relationship with the El Niño region. Considering the difference in performance between the CNN and the linear model, it indicates that the effects that the CNN captures are valuable for predicting precipitation in the central region.

Table 1: CNN Precipitation Prediction Results

R^2 Values				
	CESM2 Data (Train: 80 sims Test: 10 sims)		*Real World Data (Test Years 1990-2021)*	
California Region	***El Niño Linear Model***	***CNN***	***El Niño Linear Model***	***CNN (Fine Tuned)***
Northern	0.000	0.037	0.005	0.089
Central	0.084	0.132	0.029	0.311
Southern	0.155	0.212	0.108	0.336

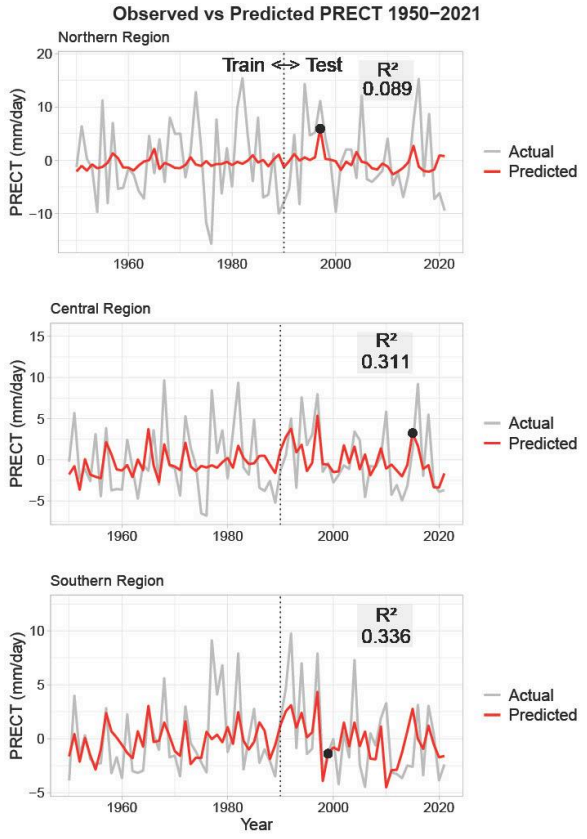

Figure 6: CNN Performance on Real World Data (training and testing)

The Southern California model performed the best, capturing 0.336 of the variability, compared to the linear El Niño model which captures 0.108 (see Table 1). Our predictions show the model's ability to capture the sign of the deviation from the precipitation mean. However, we can see in the Southern Region predicted and actual precipitation over time (Figure 6) that the model was more successful at capturing negative and low precipitation than high positive precipitation. The saliency maps for the predictions of each year's winter rainfall in the South differ greatly, with many showing the importance of the El Niño region. However, in the saliency map (see Figure 9) for our model's most accurate prediction in the year 1999, we see that the Atlantic Ocean and Indian Ocean closer to Australia were particularly important.

Our results are promising for the possible application of CNNs to predicting precipitation. Using saliency maps, we can further understand the feature importance of specific areas' sea surface temperatures.

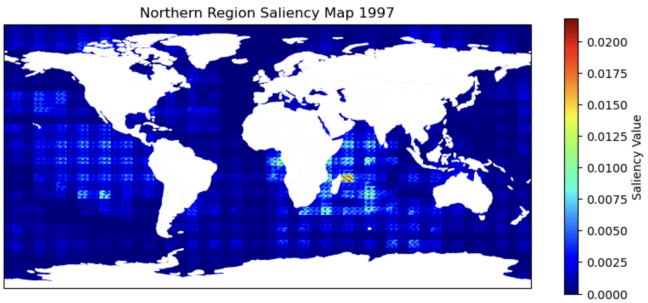

Figure 7: Saliency Map for Northern Region in 1997 (see marked point on Figure 7 for prediction)

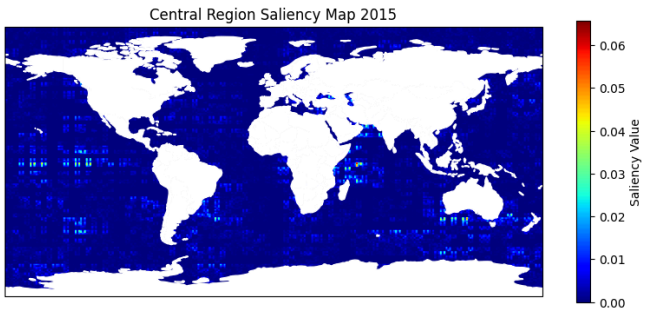

Figure 8: Saliency Map for Central Region in 2015 (see marked point on Figure 7 for prediction)

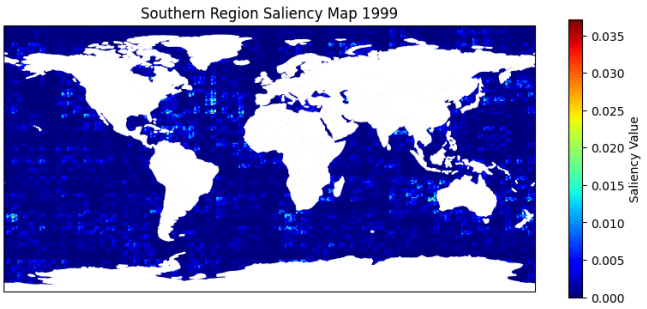

Figure 9: Saliency Map for Southern Region in 1999 (see marked point on Figure 7 for prediction)

IV. CONCLUSION

Our Convolutional Neural Networks trained on maps of global summer sea surface temperatures were able to predict northern, central, and southern California precipitation to varying degrees of success. Although both northern and central California exhibit lower predictability in existing literature, our models were able to significantly outperform a baseline linear regression in all regions by introducing non-linearity. Our paper establishes CNNs as a successful and promising solution to predicting California precipitation, and there are many future steps to consider for this approach.

Our project leveraged the CESM2 state-of-the-art simulation data for our model, resulting in 100 climate simulations and 7400 data points. However, more simulation data is accessible from sources such as the UK Earth System Modeling Project (UKESM) or the Canadian Earth System Model (CanESM5), which would allow us to increase our training sample size. Furthermore, in addition to leveraging global sea surface temperature, our model may become more successful after adding more global maps of available predictors as an input to the CNN, such as atmospheric pressure or wind fields.

Another opportunity for future research for our project would be to change or introduce more time lags within our data; we may be able to discern more information about California precipitation if we leverage the data for each predictive month of July-October or response month of November-March, rather than using the average of these seasons for each year. Additional time lags could be explored such as days or different months, although one potential concern is that introducing new predictors or time lags would greatly increase the dimensionality of the problem and the complexity of our model, or introduce unwanted interdependencies in our data that could negatively impact performance.

Beyond employing more data simulations, predictors, and time inputs, future work for this project could involve auditing the structure and hyperparameters of our existing models for each California region. The addition of more dropout or pooling layers, exploration of different activation functions, or change in the total number of model parameters could all improve the baseline success of our regional models. Hyperparameters for model training such as batch size, learning rate, patience, and the number of epochs can also be further tuned to impact model performance.

Lastly, to better understand the factors that contributed to the successes of our region models, more steps can be taken to improve Explainable AI. Saliency maps can be produced for each year of California precipitation predictions, and techniques such as principal component analysis can be applied to find the systematic source regions of predictability that most contribute to the success of our models.

REFERENCES

[1] Allen, R., Luptowitz, R. El Niño-like teleconnection increases California precipitation in response to warming. *Nat Commun* **8**, 16055 (2017). https://doi.org/10.1038/ncomms16055

[2] Gibson, P.B., Chapman, W.E., Altinok, A. et al. Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts. *Commun Earth Environ* **2**, 159 (2021). https://doi.org/10.1038/s43247-021-00225-4

[3] Liu, T., Schmitt, R. W., & Li, L. (2018). Global search for autumn-lead sea surface salinity predictors of winter precipitation in southwestern United States. *Geophysical Research Letters*, **45**(16), 8445–8454. https://doi.org/10.1029/2018gl079293

[4] Mamalakis, A., Yu, JY., Randerson, J.T. et al. A new interhemispheric teleconnection increases predictability of winter precipitation in southwestern US. *Nat Commun* **9**, 2332 (2018). https://doi.org/10.1038/s41467-018-04722-7

[5] Stevens, A., Willett, R., Mamalakis, A., Foufoula-Georgiou, E., Tejedor, A., Randerson, J. T., Smyth, P., & Wright, S. (2020). Graph-Guided Regularized Regression of Pacific Ocean Climate Variables to Increase Predictive Skill of Southwestern U.S. Winter Precipitation. *Journal of climate*, **34**(2), 737–754. https://doi.org/10.1175/jcli-d-20-0079.1