Final Figures

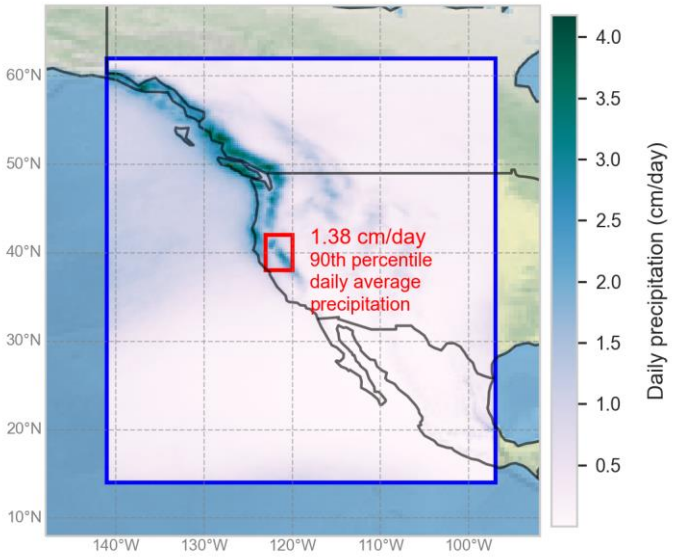

Figure 1. The Sacramento Basin study area, approximated by the red bounding box. The blue box shows the region over which spatial precipitation forecasts are considered. The map shows the 90th percentile daily precipitation magnitude over the period 2011-2019 (ERA5).

Format .png because .svg might be too big

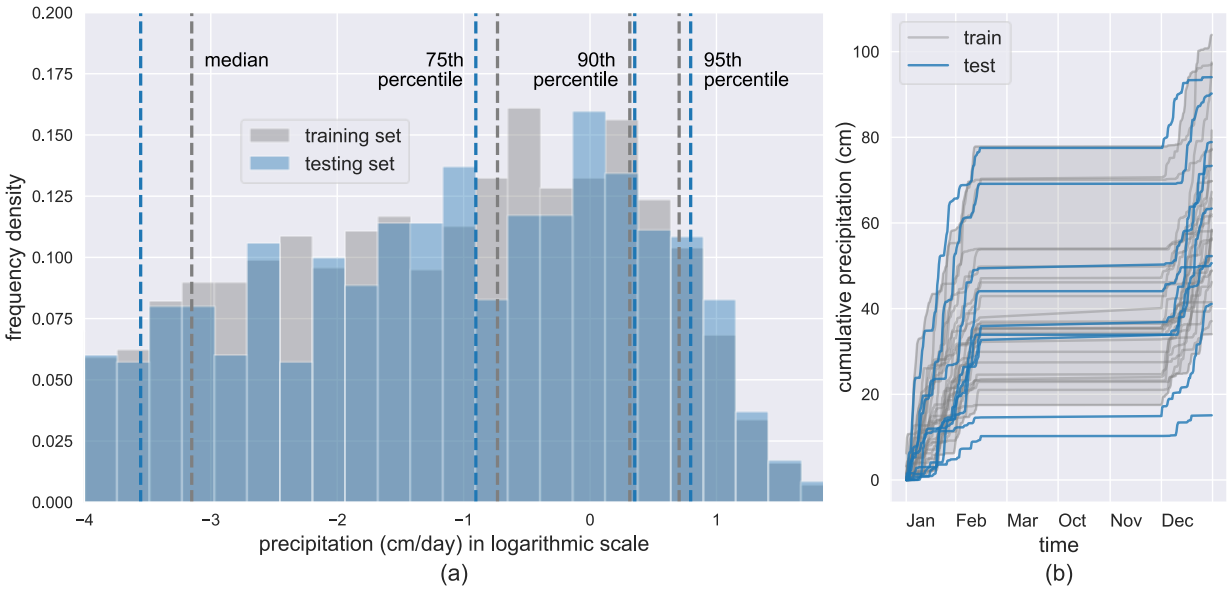

Figure 2. (a) Histogram showing the average daily precipitation over the target area during the training and test period (in logarithmic scale, with zeros omitted). (b) Cumulative daily precipitation throughout the year, showing similar seasonal patterns between the training and test sets.

Format .svg

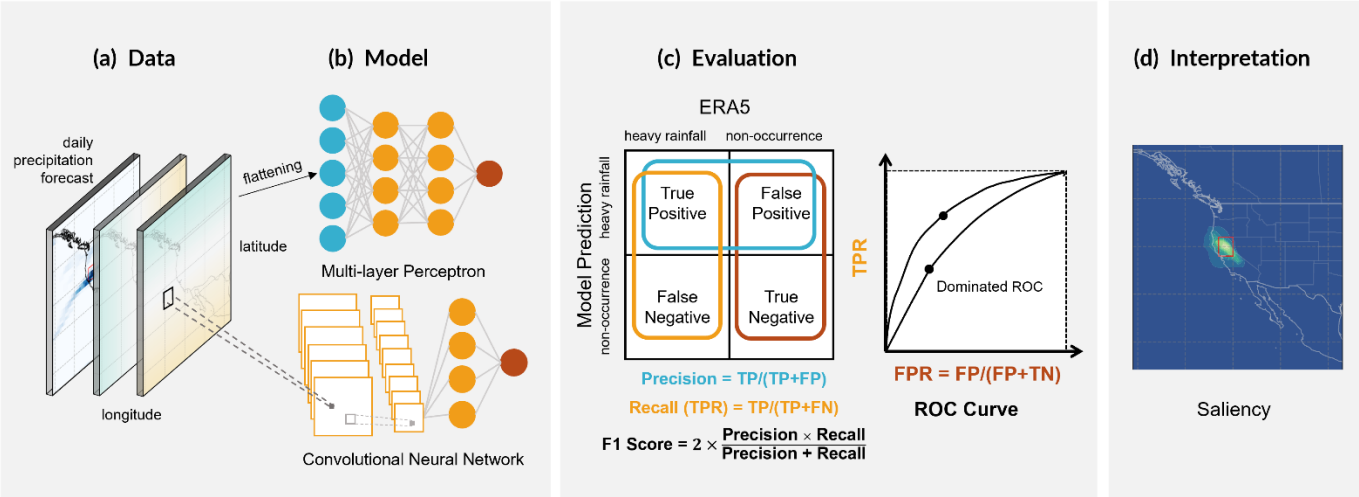

Figure 3. Experimental steps for a model defined for a specific precipitation quantile and forecast lead time.

Format .png because made in ppt

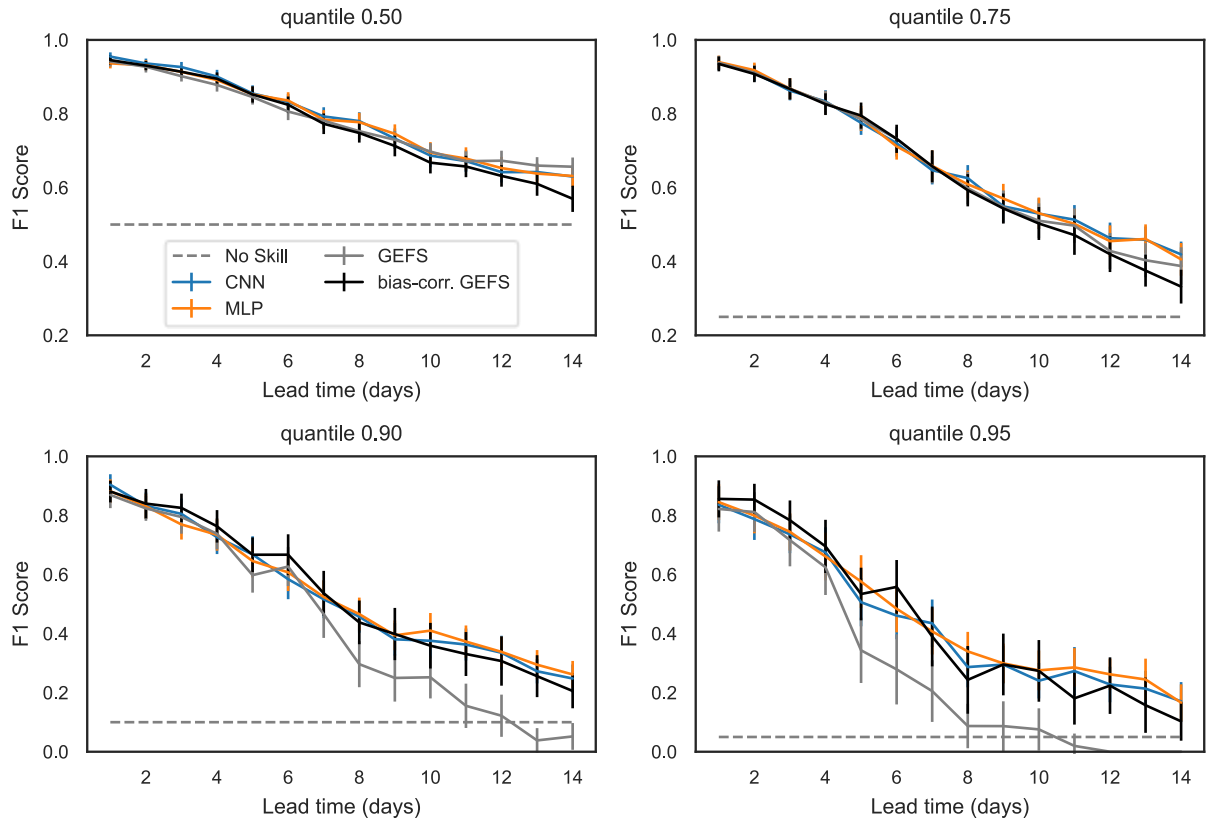

Figure 4. F-1 score comparison between CNN, MLP, GEFS, and the bias corrected GEFS benchmark for various quantiles and lead times (1-14 days).

Format .svg

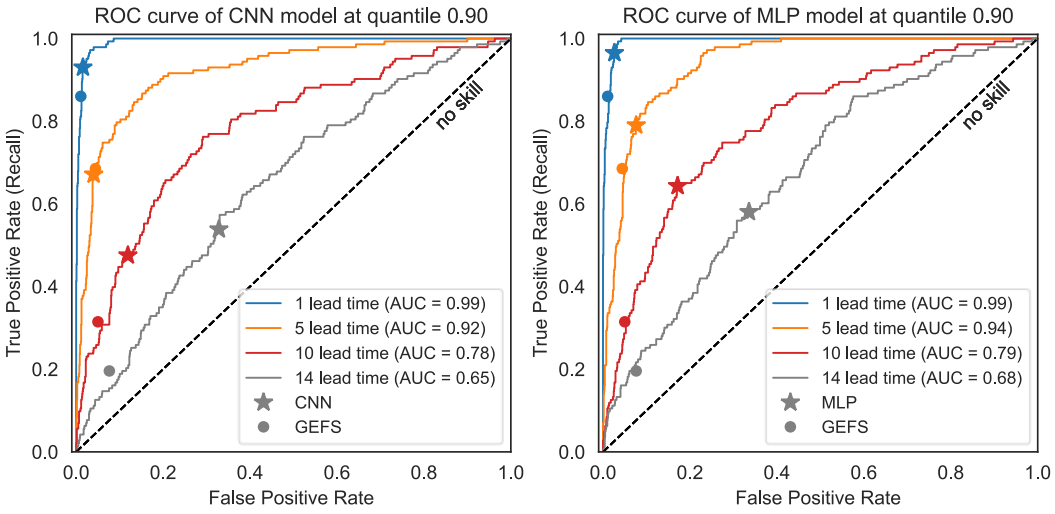

Figure 5. ROC Curve showing the tradeoff between True Positive Rate (TPR) and False Positive Rate (FPR) at different lead times for both DL models. For clarity, only lead times of 1, 5, 10 and 14 days are shown. TPR is also called the probability of detection and FPR is called the probability of false alarm.

Format .svg

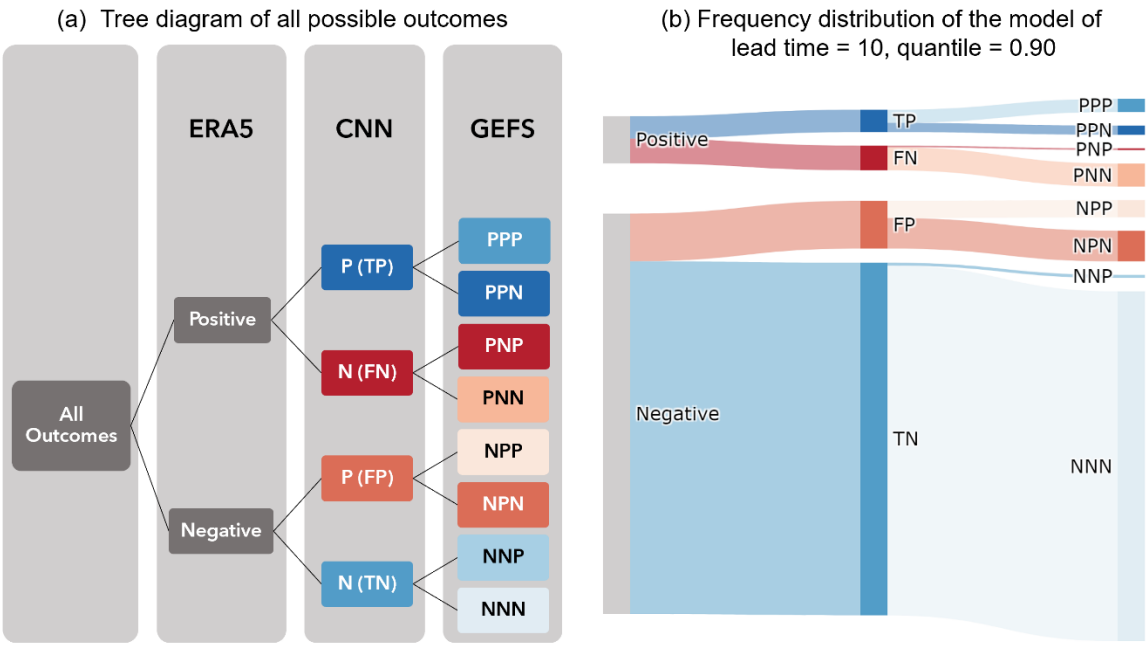

Figure 6. (a) 8 outcome scenarios incorporating the performance of benchmark. Blue indicates preferred outcomes, while red denotes undesirable outcomes. (b) Fraction of days in each scenario over the testing period (2011-2019) for one specific example (lead time 10 days, quantile 0.9). The first two layers denote the commonly used confusion matrix, while the third incorporates the bias-corrected GEFS benchmark.

Format .png because made in ppt

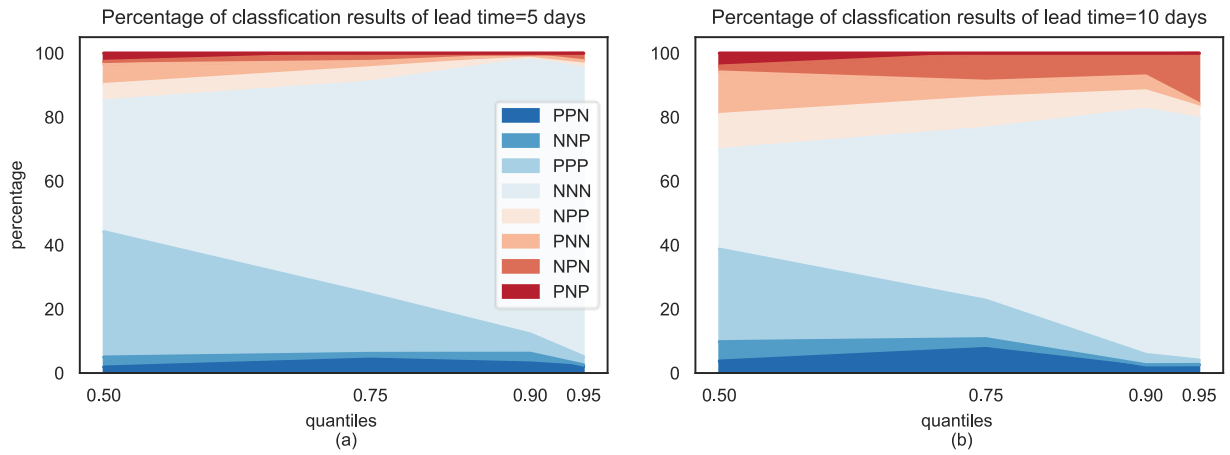

Figure 7. Tradeoff between forecast corrections and misclassifications at lead times of (a) 5 days, (b) 10 days. The rate of misclassifications for both models (PNN and NPP) increases with longer lead times and more extreme quantiles.

Format .svg

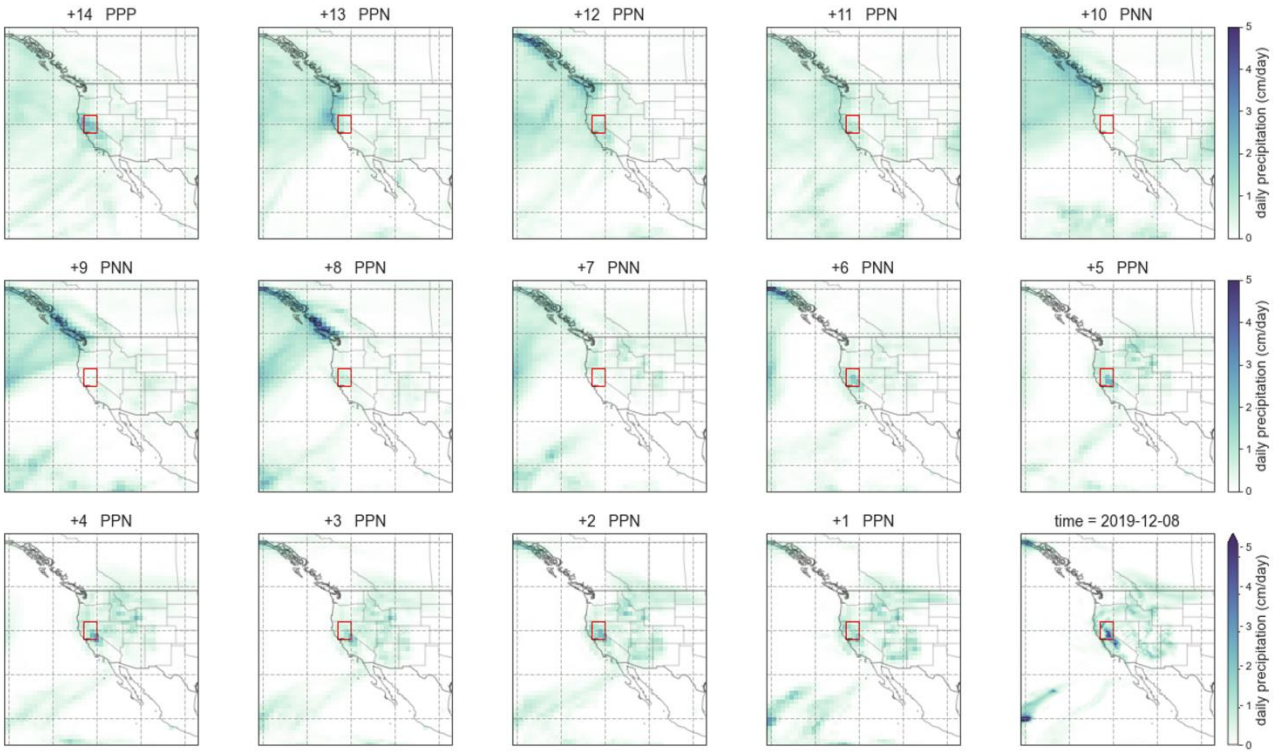

Figure 8. 14-day GEFS forecast leading up to the AR event on December 8, 2019. This was a heavy precipitation event for which the GEFS forecast is often corrected by the CNN model, accounting for both spatial error (e.g., at lead time 8 days) and magnitude error (e.g. at lead time 1-5 days).

Format .png

original svg is too big, also edited in ppt.

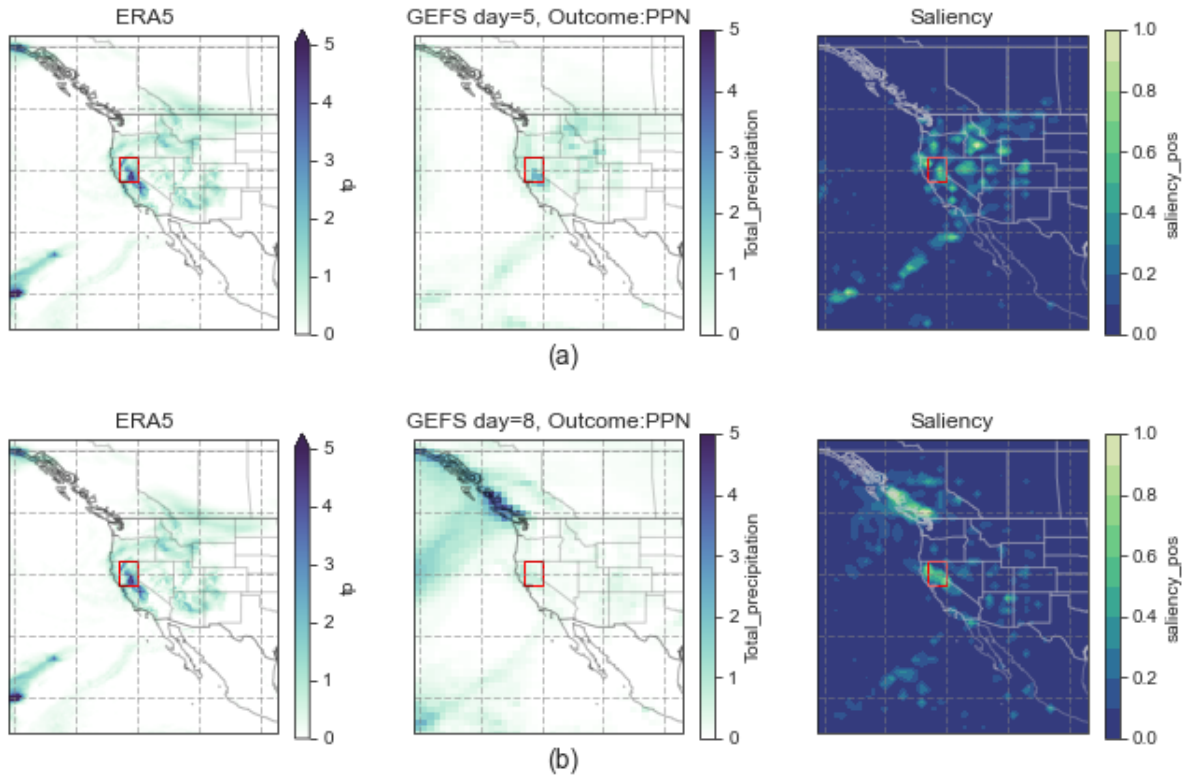

Figure 9. Example of individual forecast days leading up to the AR event in Figure 8. (a) Correction of spatial error at lead time of 8 days; (b) Correction of magnitude error at lead time of 5 days. The saliency maps indicate the attention of the CNN classifier.

Format .png because original svg is too big,

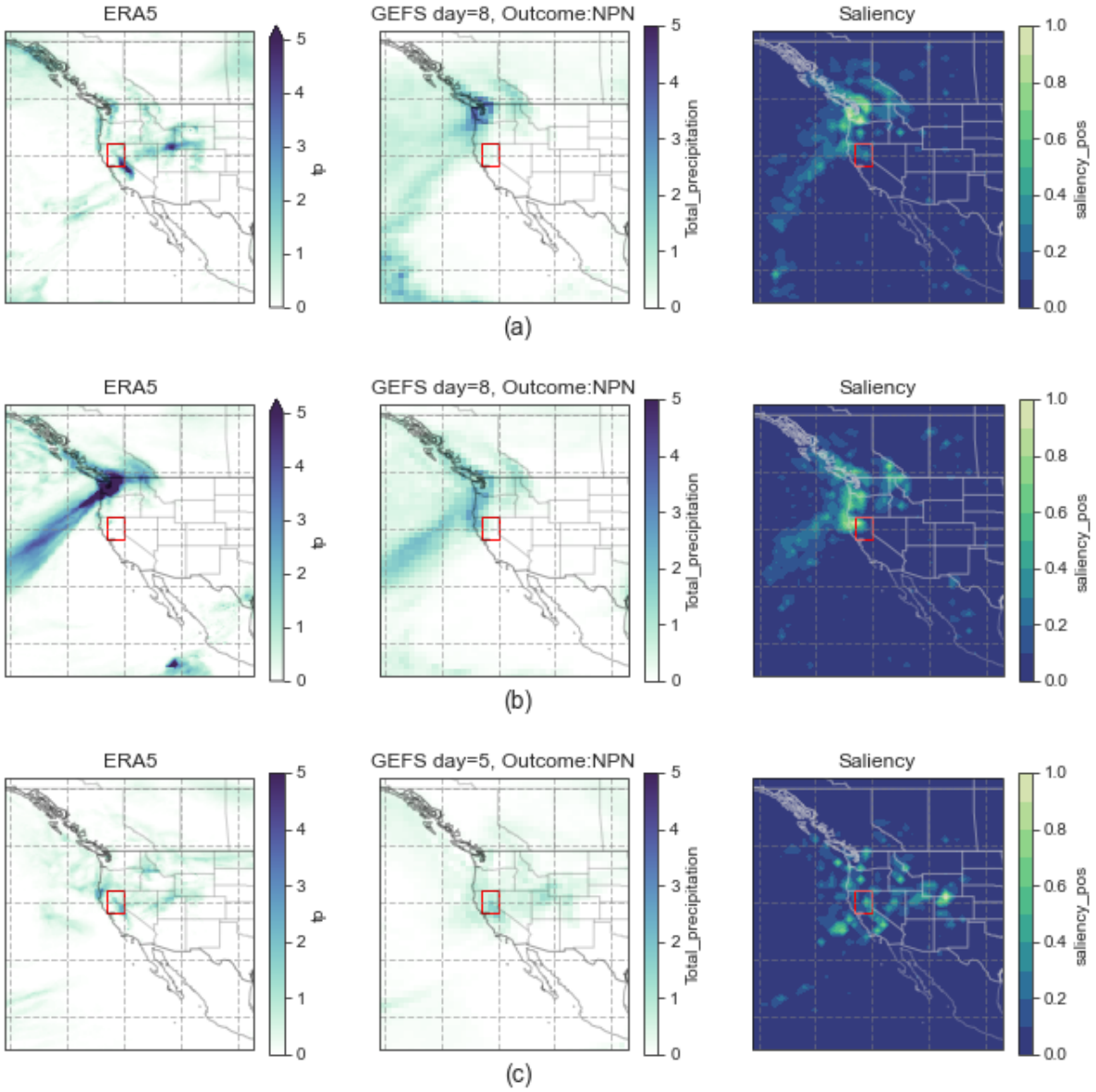

Figure 10. Examples of false positives in which the CNN model attempts to correct the GEFS forecast.

(a) November 17, 2017, (b) December 20, 2019, (c) February 22, 2017.

Format .png because original svg is too big,

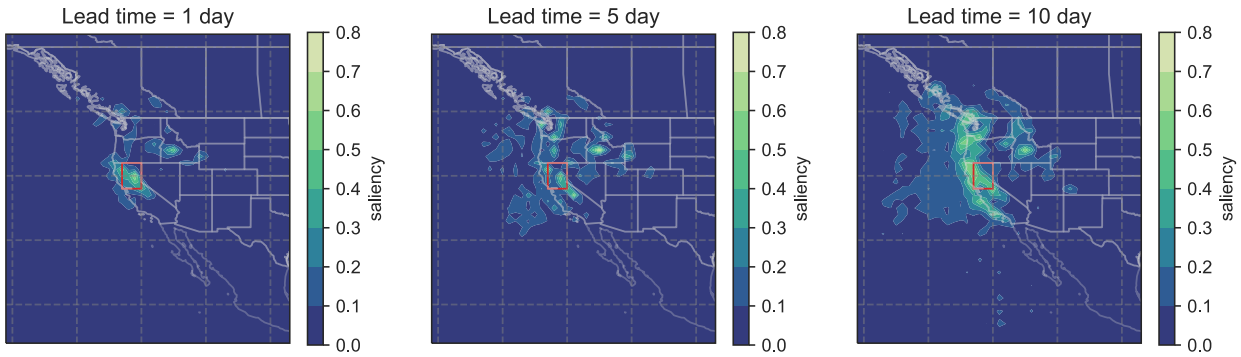

Figure 11. Average saliency maps for CNN models of precipitation quantile 0.90. As the lead time increases, the attention of the CNN model expands further away from the target region (red box).

Format .svg

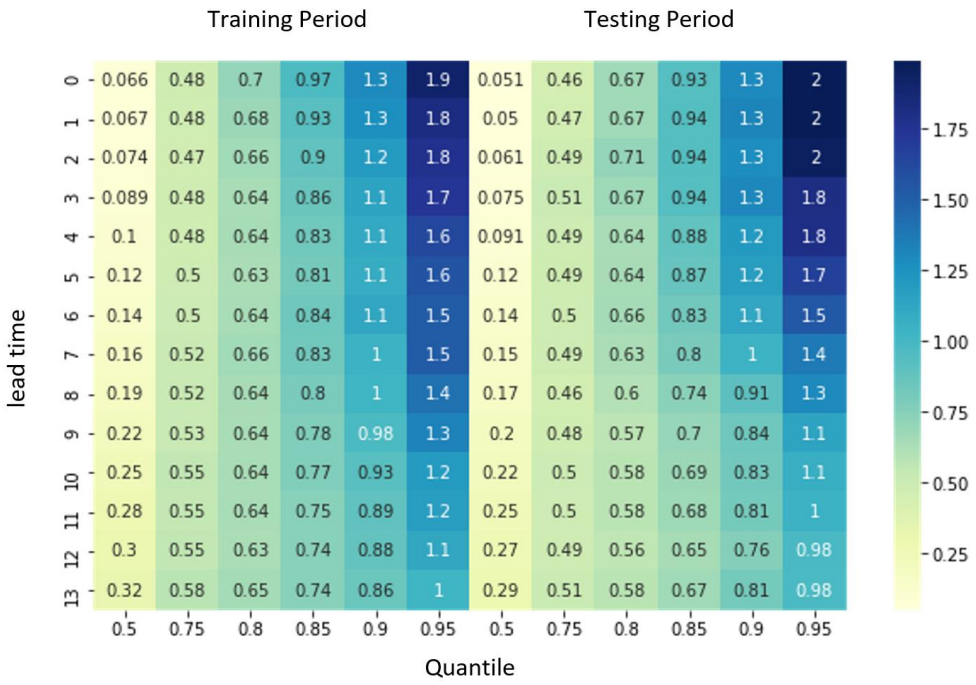

Figure 12. Quantile values for train period (1985-2010) and test period (2011-2019) displays a shift in quantile values of GEFS forecasts along with an increase of lead time. The quantile values for the entire period are used to generate the bias corrected benchmark.

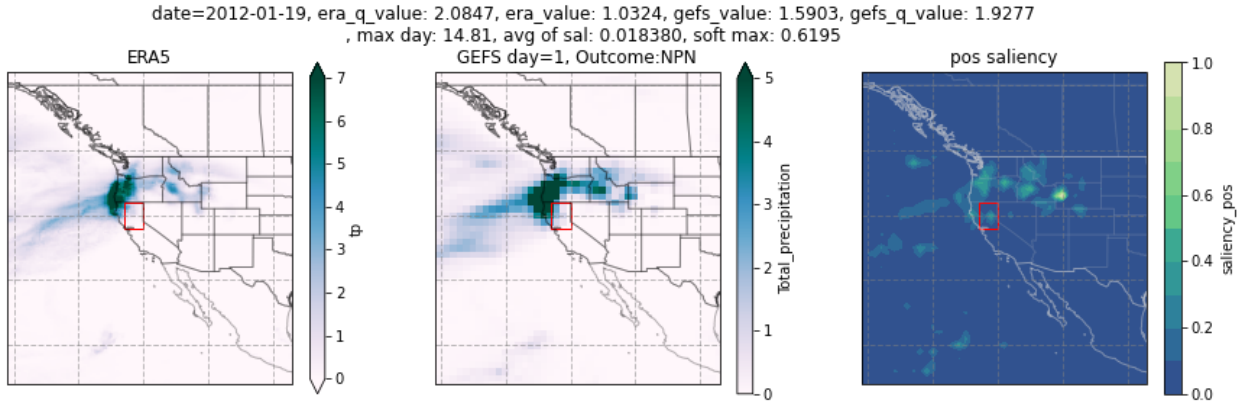

Figure 13. A False Positive instance showing that the CNN model of lead time =1 and quantile = 0.90 did not precisely learn the rectangular spatial boundary of Sacramento River Basin

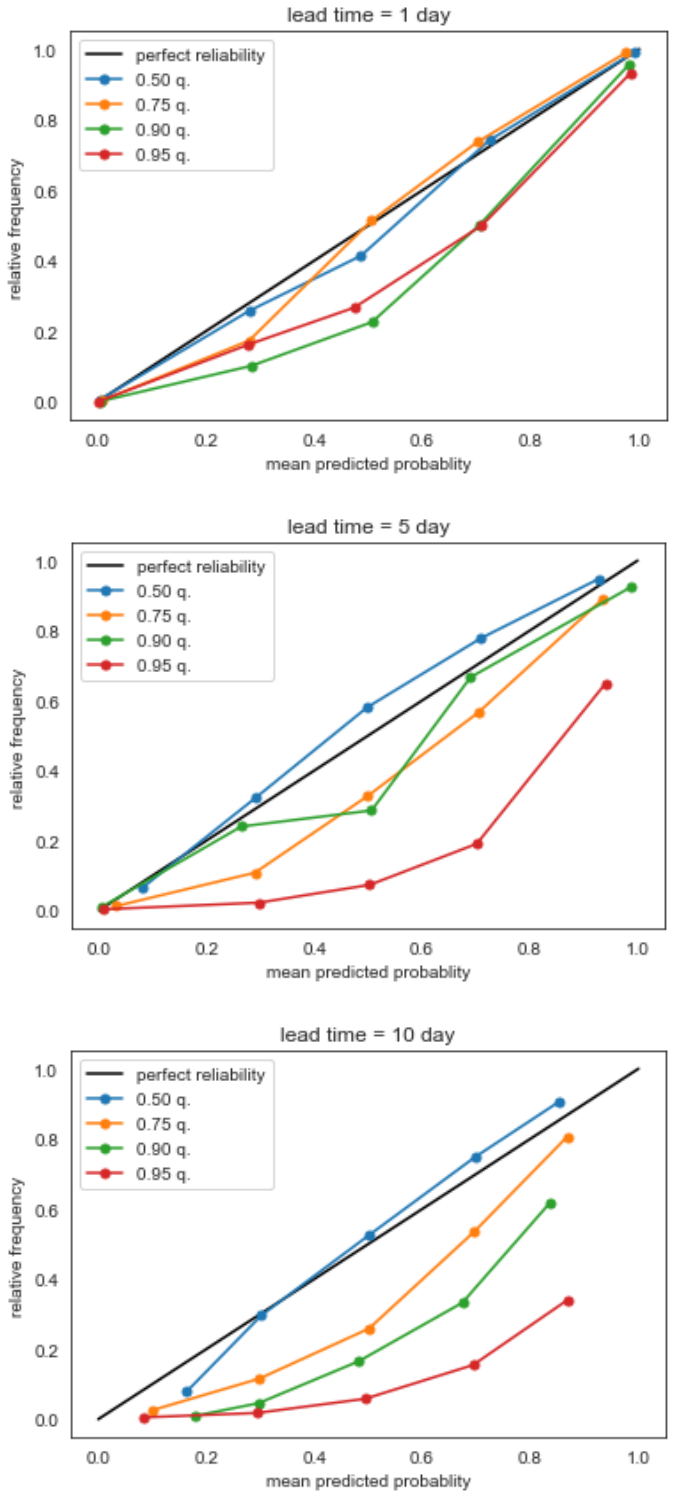

Figure 14. Reliability Diagram of CNN models for 1,5 and 10 day lead time respectively.

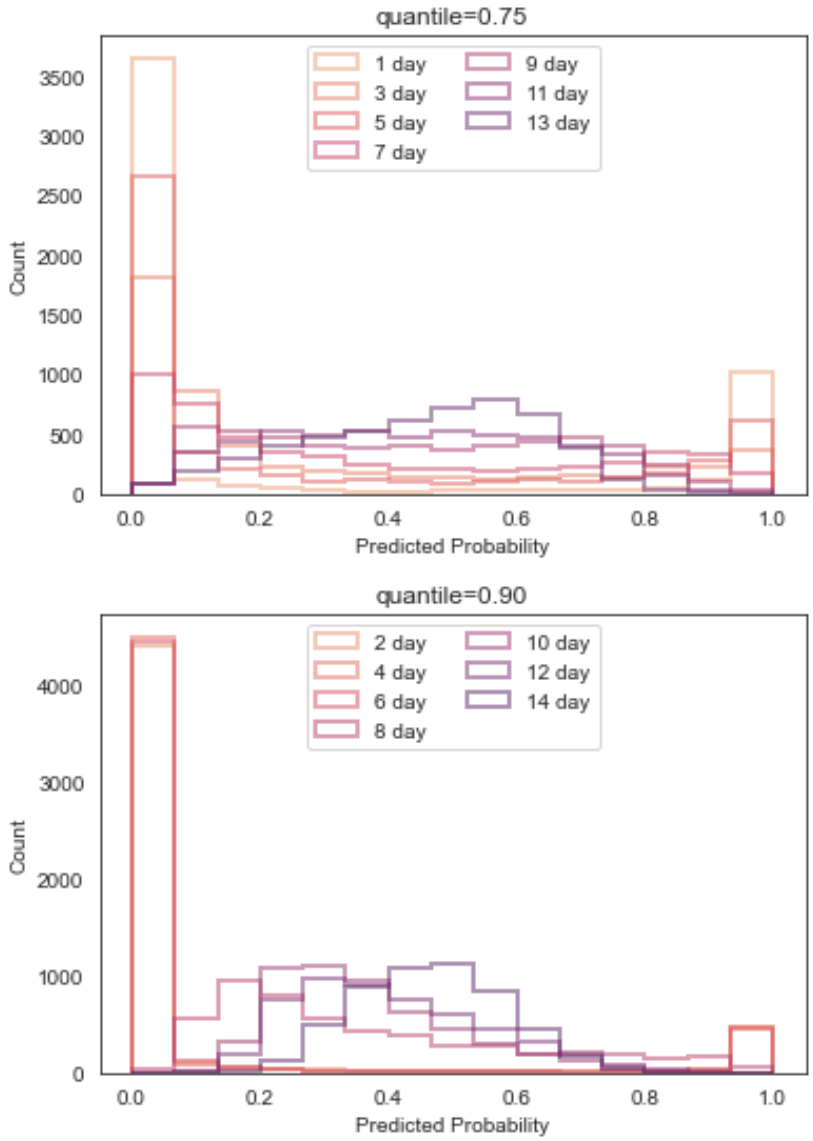

Figure 15. Histogram of CNN models across different lead times of quantile 0.75 and 0.90. The probability distribution shifts from bimodal to unimodal distribution as lead time increases, suggesting lower confidence facing a more difficult prediction problem. a) CNN model 0.75 quantile, b) CNN model 0.90 quantile.

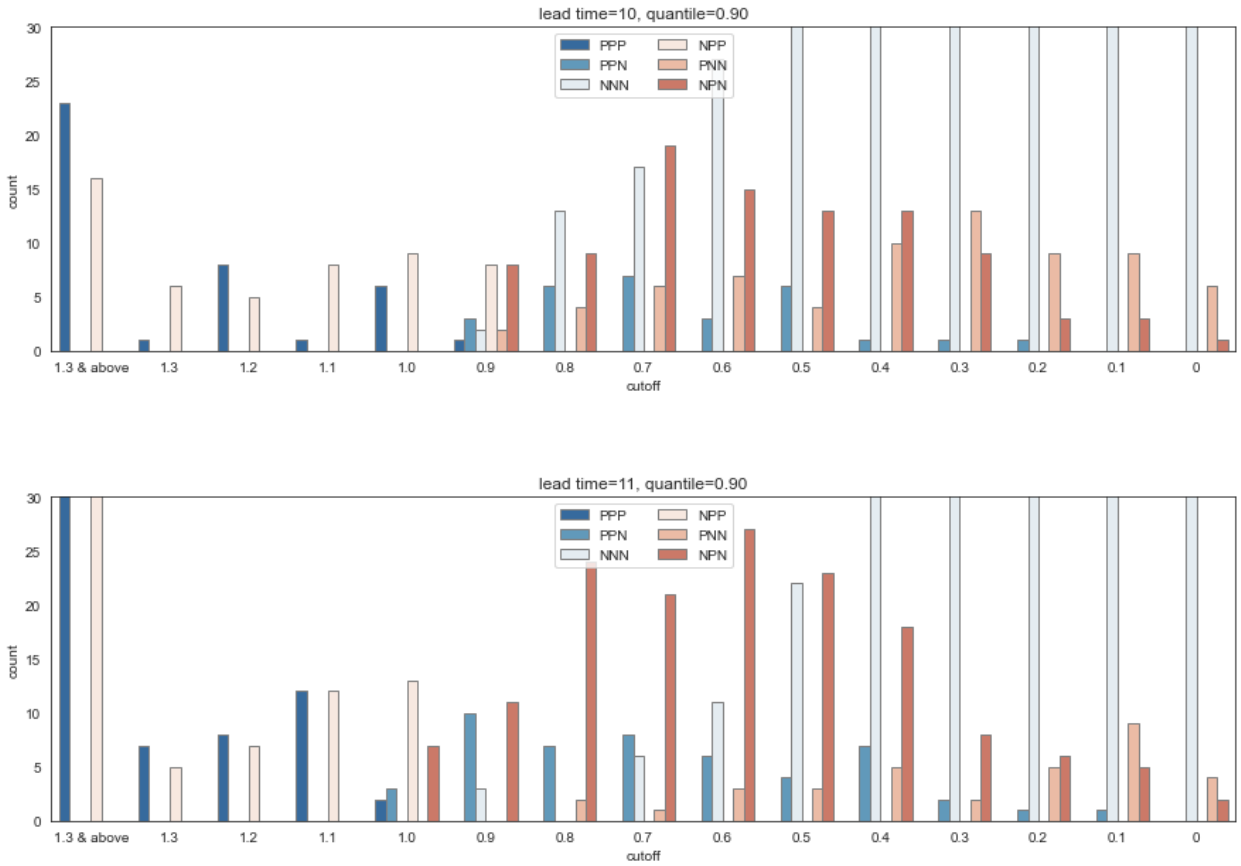

Figure 16. Comparison of NPN (false alarms) of lead time = 10 and lead time = 9 for CNN model.