Enhancing Deterministic Freezing Level Predictions in the Northern Sierra Nevada Through Deep Neural Networks

Vesta Afzali Gorooh, Agniv Sengupta, Shawn Roj, Rachel Weihs, Brian Kawzenuk,

Luca Delle Monache, F. Martin Ralph [a]

Center for Western Weather and Water Extremes, Scripps Institution of Oceanography

University of California, San Diego

Corresponding author: vafzaligorooh@ucsd.edu

1

ABSTRACT

Accurate prediction of the freezing level (FZL) is essential for hydrometeorological forecasting systems and precipitation phase estimation, and it influences runoff generation and reservoir management decisions. In this study, we develop a deep learning–based postprocessing framework using the Unet convolutional neural network (CNN) architecture to refine the FZL forecasts from the West Weather Research and Forecasting (West-WRF) model. The proposed framework leverages reforecast data from West-WRF and FZL estimates from the California-Nevada River Forecast Center (CNRFC) to train a deterministic Unet model over the Yuba-Feather watershed, a hydrologically critical basin in northern California. We introduce two variants of our model, Unet–Log and Unet–GMM, which utilize the logarithm of the hyperbolic cosine of Error and Gaussian Mixture Model loss functions, respectively, to enhance FZL forecast accuracy beyond an RMSE-based benchmark. Results indicate that the Unet–based postprocessing framework significantly improves FZL forecast skill across diverse atmospheric conditions and complex topography. Compared to the raw West-WRF output, our model achieves reductions in RMSE of up to 25% and increases the forecast–observation correlation by about 10% over the Yuba-Feather watershed. Furthermore, it effectively captures the spatiotemporal variability of the FZL across different elevations, mitigating systematic biases inherent in the West-WRF model. This novel deep learning–based postprocessing approach demonstrates a promising pathway for integrating machine learning into hydrometeorological forecasting and decision support within the Forecast Informed Reservoir Operations (FIRO) framework.

1. Introduction

Accurate estimates of the atmospheric freezing level (FZL)—defined in this study as the altitude above mean sea level (MSL) at which the atmospheric temperature reaches 0°C—is crucial for a wide range of high-impact applications in complex terrain regions, including precipitation type discrimination, hazard mitigation efforts, and water resource management. In the Northern Sierra Nevada Mountains, where complex topography and rapidly changing atmospheric conditions interact, landfalling atmospheric river events (AR; e.g., Zhu and Newell 1994, 1998; Ralph et al. 2004, 2005) play an integral role in shaping the water year (Neiman et al. 2008; Dettinger 2011; Dettinger and Cayan 2014). Dynamical models such as the Weather Research and Forecasting (WRF) model often produce FZL forecasts with significant biases and uncertainties. Observations and models have shown the complexity and

variability of the FZL in this mid-latitude mountainous region (Sumargo et al. 2020; Henn et al. 2020; Minder and Kingsmill 2013; Cannon et al. 2017). Dynamical model forecast errors arise from multiple sources, including initial conditions, boundary conditions, and structural model deficiencies (Delle Monache et al. 2013) . The chaotic nature of atmospheric dynamics further amplifies these errors over time, limiting forecast skill and operational reliability. These limitations impair the accuracy of operational forecasts and exacerbate risks associated with extreme precipitation and unexpected icing or runoff events.

Recent studies have highlighted the nonlinear relationship between the FZL and hydrological responses, particularly runoff variability. Qin et al. (2019) analyzed the FZL and summer runoff in the Hotan River Basin, China, using nonlinear decomposition techniques and hybrid prediction models. The study revealed significant inter-annual and inter-decadal FZL cycles, emphasizing the importance of accounting for nonlinear interactions in hydrological modeling. These results underscore the critical role of advanced data-driven approaches in improving FZL-based hydrological forecasting, which has implications for flood risk assessment and water resource management. Furthermore, Sumargo et al. (2020) investigated the consequences of FZL forecasts on reservoir flood control storage in California, specifically for Lake Oroville and New Bullards Bar Reservoirs, demonstrating that FZL forecast errors can lead to significant misrepresentations of storage, consuming 2–10% of flood pool volume for every inch of precipitation, depending on local conditions. Their findings underline the sensitivity of the Feather and North Fork Yuba River watersheds to FZL variability and emphasize the need for forecasting improvements to enhance flood risk management and reservoir operations.

Postprocessing techniques have proven beneficial in related atmospheric forecasting domains. For example, bias correction methods, Kalman filters (e.g., Delle Monache et al. 2011), and multi-linear regression-based postprocessing techniques, referred to as Model Output Statistics (MOS; Glahn and Lowry 1962)–and their extensions, such as Ensemble MOS (EMOS), have been widely employed to systematically improve the accuracy in numerical weather prediction (NWP) forecasts by establishing statistical relationships between raw model outputs and observations (e.g., Wilks and Hamill 2007; Taillardat et al. 2016). However, these methods are inherently limited in their ability to capture the complex, highly non-linear interactions governing the vertical thermal structure of the atmosphere and,

3

as a result, often struggle to accurately model rapid changes in the FZL, inversions, and other key thermodynamic features (Vannitsem et al. 2021).

Deep learning postprocessing approaches offer a promising alternative to traditional statistical methods. Recent studies have successfully enhanced both deterministic and probabilistic short-term forecasts of near-surface temperature (e.g., Alerskans et al. 2022; Cho et al. 2022), precipitation (e.g., Ghazvinian et al. 2021, 2022; Hu et al. 2023; Ghazvinian et al. 2024; Badrinath et al. 2023), sea ice (Palerme et al. 2024), wind speed (Sun et al. 2024) and integrated vapor transport (IVT; (Chapman et al. 2019, 2022). These methods, such as convolutional neural networks (CNNs) and Unet architectures, leverage large, gridded datasets and customized loss functions to capture fine-scale spatial features and reduce forecast errors, demonstrating a potential to overcome conventional postprocessing limitations (Pan et al. 2019; Zhang et al. 2023). For instance, Li et al. (2022) used a LeNet-type CNN to enhance precipitation forecasts by predicting precipitation distribution parameters from multiple atmospheric predictors such as moisture and circulation features. Although this approach improved spatial representation, its use of coarse-resolution NWP inputs limited its ability to capture finer-scale structures. Similarly, Chapman et al. (2019, 2022) showed that CNN-based postprocessing of atmospheric river forecasts and IVT predictions—using outputs from the National Weather Service's Global Forecasting System (GFS) and a WRF version designed for the prediction of extreme events associated with atmospheric rivers over the Western US, called West-WRF (Delle Monache et al. 2024)—reduced forecast errors and enhanced spatial correlation with observations. Badrinath et al. (2022) further demonstrated that Unet-based models consistently reduced systematic and random errors while preserving spatial consistency compared to traditional methods. In addition, studies by Ghazvinian et al. (2021) and Hu et al. (2023) underscored the benefits of customized loss functions in deep neural network frameworks for improving forecast skills. These findings are particularly relevant for FZL forecasting, where even minor errors can lead to significant discrepancies in predicting precipitation type and associated hazards. Advanced postprocessing methods that correct systematic biases and capture fine-scale variations in atmospheric thermal structure are essential. By incorporating techniques such as deep learning with Unet architectures and customized loss functions, it can substantially improve FZL forecasts.

Inspired by these insights, this study examines whether an end-to-end Unet-based postprocessing framework can effectively exploit high-dimensional data to uncover the nonlinear relationships between atmospheric variables that influence FZL dynamics. Our approach integrates raw dynamical model outputs from the regional West-WRF model with geographical information to reduce systematic biases and random errors. As a proof of concept, we focus on the Yuba-Feather watershed—a region with a long history of catastrophic floods and a primary basin under the Forecast-Informed Reservoir Operations initiative (FIRO; Talbot et al. 2019; Jasperse et al. 2015, 2020). We hypothesize that leveraging the nonlinear feature extraction capabilities of the Unet architecture will enhance the representation of complex spatial and temporal interactions among orographic, thermodynamic, and dynamic influences, thereby improving FZL forecast accuracy.

The remainder of this paper is organized as follows: section 2 describes the data sources utilized in our study, while section 3 details the proposed postprocessing methodology over the study area. Section 4 presents the model validation and performance assessment against observational datasets.

2. Data Sources

a. Dynamical Model Forecasts

In this study, FZL forecasts from the West-WRF reforecast (Cobb et al. 2023) were used for the postprocessing model development. The reforecast is a 33-year-long dataset over the Western U.S. with diverse atmospheric and hydrological conditions. It provides insights into representative high-resolution atmospheric dynamics during the months of December through March. This dataset is a valuable source of information for data-driven algorithm development, such as deep learning postprocessing methodologies and in-depth process-based studies (Cobb et al., 2023). The West-WRF reforecast is based on WRF version 4.0.1 and is initialized daily at 0000 UTC using the Global Ensemble Forecast System Version 10, which refines the global model's coarse 0.5-degree resolution down to high-resolution 9 km and 3 km grids. The reforecast employs 60 vertical levels with an adaptive time-stepping approach, optimizing numerical stability across varying atmospheric conditions. The 3-km domain used in this study provides high-resolution forecasts for detailed analysis for 120 hours, while the 9-km domain offers extended lead times of up to 168 hours. The forecast files were generated every 3 hours, allowing for comprehensive temporal resolution in the

training dataset. Nine winter seasons (December 2010 – March 2019) were used for the postprocessing model training and evaluation, capturing the period when FZL fluctuations are most critical. Since FZL is not directly provided in the reforecast product, the first instance of a 0°C crossing is identified for each grid cell, beginning at the top of the model. If no 0°C crossing is found—indicating that the surface remains below freezing—the FZL is set to the model terrain height.

b. Ground Truth / Reference

The California-Nevada River Forecast Center (CNRFC) observed FZL was used as ground truth (reference dataset) in this study. The gridded CNRFC dataset represents instantaneous FZL heights at six hourly resolutions (0000, 0600, 1200, and 1800 UTC). Before 1 Oct 2020, the FZL grid was created by interpolating sixty-eight points from the GFS model (D. Kozlowski, personal communication, June 6, 2023) onto the 4-km HRAP (Hydrologic Rainfall Analysis Project) grid that matches the CNRFC's other operational products.

3. Methodology and Experimental Design

a. Preprocessing West-WRF reforecast and CNRFC data:

The West-WRF reforecast dataset and the CNRFC FZL analysis files were harmonized through a series of preprocessing steps to facilitate a robust postprocessing analysis. Given that the West-WRF reforecasts are at a native 3-km resolution, while the CNRFC FZL analyses are available on a 4-km HRAP grid, a careful regridding procedure was essential to ensure spatial consistency.

1. *Th*e CNRFC files, which serve as the ground truth for FZL analyses, were retained without modification. The HRRR Digital Elevation Model (DEM) was rescaled to the 4-km CNRFC HRAP grid.

2. Any CNRFC FZL value that fell below the corresponding elevation from the 4-km DEM was set equal to the DEM elevation, ensuring that freezing level values did not drop below the physical terrain.

3. Within each West-WRF reforecast file, the first occurrence of a 0°C crossing was identified for each grid cell, starting from the top of the model domain. In cases where no

6

0°C crossing was observed, indicating that the entire atmospheric column remained below freezing, the FZL was assigned the value of the terrain height.

4. The computed 0°C crossing elevations from the West-WRF reforecast (originally on the 3-km grid) were gridded to align with the 4-km HRAP grid of the CNRFC FZL analysis files.

These preprocessing steps ensured consistency between the West-WRF and CNRFC datasets, supporting the development of the deep learning postprocessing framework described in Section 3.2.

b. Unet Framework

The CNN with Unet architecture, introduced by Ronneberger et al. (2015), has been widely used and consists of a contracting path to capture context and a symmetric expanding path that enables precise localization. Figure 1 shows a general Unet adapted to process West-WRF model outputs and predict the FZL in an effort to improve accuracy. The encoder systematically reduces the spatial dimensions of the input data (West-WRF FZL) while increasing the number of feature channels. This is achieved through convolutional layers with 3×3 kernels applied sequentially. Each convolutional layer is followed by a Leaky rectified linear unit (LeakyReLU) activation function to introduce non-linearity. After every convolution set, a 2×2 max-pooling operation is utilized to downsample the feature maps effectively and halve the spatial dimensions, allowing the model to capture hierarchical features at multiple scales.

The decoder reconstructs the spatial dimensions by upsampling the feature maps. Upsampling is performed using transposed convolutions followed by standard convolutions. At each decoder stage, the upsampled feature maps are concatenated with the corresponding feature maps from the encoder via skip connections. This fusion combines the high-resolution spatial information from the encoder with the higher-level features from the decoder, enhancing the network's ability to produce precise outputs. The skip connection mechanism preserves spatial details and facilitates the flow of gradients during training, leading to improved convergence and performance. The final layer of the Unet employs a 1×1 convolution to map the multi-channel feature maps to the desired output (here, postprocessed FZL). This study implemented the Unet neural networks to refine the FZL from West-WRF model outputs. The input data is formatted similarly to image (latitude, longitude) data,

7

allowing the network to leverage spatial patterns and correlations inherent in atmospheric phenomena.
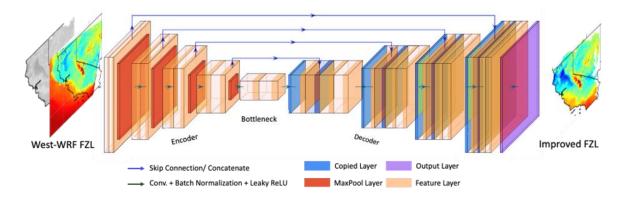

Figure 1. Schematic of the Unet architecture for freezing-level postprocessing.

c. *Loss function and model training*

First, we train a Unet model with root mean squared error (RMSE) loss function (Unet-RMSE), hereafter used as a benchmark for the two proposed Unet models that use customized loss functions. The two customized loss functions are: (1) the Log-Cosh error function, which is a smooth approximation of mean absolute error (MAE) and combines the advantages of mean square error (MSE) and MAE, and (2) a Gaussian Mixture Model (GMM) loss function that considers the distribution of the prediction. The details of each of these loss functions are provided below.

LOG-COSH ERROR LOSS FUNCTION

The Log-Cosh loss function is defined as the logarithm of the hyperbolic cosine of the prediction error (Jadon et al. 2024). It provides a robust alternative to traditional loss functions by smoothly approximating MAE and combining the benefits of MSE and MAE. This effectively handles both small and large errors, which is particularly useful in meteorological data, where outliers can significantly affect model performance.

The loss function is defined as

$$Loss = \frac{1}{N}\sum_{i=1}^{N} log(cosh(\hat{y}_i - y_i)) \tag{1}$$

Where N is the number of samples, y_i is the true value for the ith sample and $\hat{y}_i$ is the predicted value.

GAUSSIAN MIXTURE MODEL LOSS FUNCTION

File generated with AMS Word template 2.0

The GMM loss function models the prediction error as a mixture of two Gaussian distributions, allowing the model to capture complex error distributions. The probability density function $F(x)$ is defined as:

$$F(x) = \pi_1 . N(x, \mu_1, \sigma_1^2) + \pi_2 . N(x, \mu_2, \sigma_2^2) \tag{2}$$

Where π_1 and π_2 are the mixing coefficients such that $\pi_1 + \pi_2 = 1$ and $N(x, \mu_i, \sigma_i^2)$ is the Gaussian distribution of x:

$$N(x, \mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}} exp(-\frac{(x-\mu_i)^2}{2\sigma_i^2}) \tag{3}$$

The loss function is then defined as the negative log-likelihood:

$$Loss = -\frac{1}{N} \sum_{i=1}^{N} log(F(x_i)) \tag{4}$$

The Unet models are trained separately using the Log-Cosh and GMM loss functions (hereafter Unet-log and Unet-GMM, respectively). We adjusted the number of layers and feature channels in the Unet architecture to balance computational efficiency with the complexity required to model the FZL accurately. Additionally, we trained different models for each lead time to better capture temporal dynamics. Training involves optimizing the network parameters to minimize loss functions using stochastic gradient descent with appropriate learning rates and regularization techniques. The datasets include West-WRF model outputs (as input to the Unet) and observed CNRFC FZL measurements (as ground truth). The training period includes two time intervals: from December 2010 to March 2015 and December 2017 to March 2019. During the validation period, a one-year cross-validation approach was employed to optimize the model's hyperparameters effectively. The test period, which was completely unseen by the model during training and validation, spanned from December 2015 to March 2017. This independent test period covers one dry (from 1 December 2015 to 31 March 2016) and one wet (from 1 December 2016 to 31 March 2017) winter season, allowing for the evaluation of the model's performance under varying climatic conditions. The 2016–17 season brought record-breaking precipitation across California, while the 2015–16 season remained unexpectedly dry despite a strong El Niño event, highlighting the role of internal atmospheric variability in modulating precipitation (Kumar and Chen 2017; Luna-Niño et al. 2025; Siler et al. 2017) Appendix A provides maps of accumulated precipitation for both winters, as well as a difference map to emphasize the spatial variability between the two seasons.

File generated with AMS Word template 2.0

Figure 2 shows the study domain encompassing the Yuba-Feather River watershed, with the topographic map including the North Fork Feather, East Branch North Fork Feather, Middle Fork Feather, and Upper Yuba basins. The Yuba River, originating in the Sierra Nevada, is one of California's historic rivers, with its main tributaries being the North, Middle, and South Yuba Rivers. The Feather River, the largest in the Sierra Nevada, begins in the northern Sierra Valley and flows into the Sacramento River. This region is of particular hydrological importance due to its complex topography, orographically enhanced precipitation, and frequent interactions with ARs. This watershed, with hydropower and flood-control reservoirs, requires accurate FZL predictions to improve flood mitigation strategies and optimize water storage efficiency. Given its highly variable precipitation regimes and sensitivity to FZL fluctuations, this region provides an ideal case study for evaluating machine learning-based postprocessing techniques aimed at reducing FZL forecast biases.
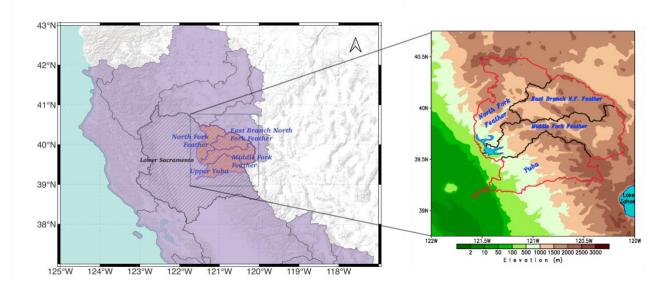

Figure 2. The study domain and the elevation map of the Yuba Feather River watershed in the inset.

4. Results and Discussions

In this section, we present the results of the Unet-based postprocessing models over two winter seasons independent of the training period: a wet year (2016-17) and a relatively dry year (2015-16). Since the reference data, CNRFC, and West-WRF (hereafter WRF) are deterministic, we evaluate the Unet-GMM model using the expected value of the predicted

10

distribution, as defined in Equation (2). We begin by presenting the time series of FZL forecast performance, followed by the spatial distribution of skill metrics comparing the original WRF forecasts and the postprocessed forecasts relative to CNRFC observations across different lead times. Finally, we assess the models' skill across varying terrain elevations to understand their performance in complex topography.

a. *Performance of freezing level forecasts postprocessed with deep learning*

Figures 3 and 4 illustrate the performance of the Unet postprocessing models (Unet-RMSE, Unet-log, Unet-GMM) as a function of lead time against CNRFC observational target over the Yuba Feather watershed for the winters of 2016–17 and 2015–16, respectively. The results are computed using winter season samples aggregated over grid points covering the study area, offering insight into the temporal evolution of forecast skill. All three Unet models exhibit noticeable improvements in FZL forecasts in both the Pearson correlation coefficient (hereafter correlation) and root mean square error (RMSE) when compared to the raw WRF forecasts during the relatively wet 2016–17 season (Figure 3). Definitions of these verification metrics are provided in Appendix B. The Unet-based methods (blue, red, and grey lines) consistently achieve higher correlation and lower RMSE values than the baseline forecasts (black line), indicating improved forecast accuracy. During the 2016–17 winter, the average RMSE across all lead times decreased from approximately 370 m in the raw WRF forecasts to 360 m with Unet-RMSE and further to 320 m with both Unet-GMM and Unet-log. Similarly, the correlation improved from an average of 0.85 in WRF to 0.88 with Unet-RMSE and further to 0.90 with the Unet-GMM and Unet-log models. These results indicate that while Unet-RMSE effectively reduces errors compared to the raw forecasts, the probabilistic models (Unet-GMM and Unet-log) provide additional gains in both accuracy and consistency. Depending on the lead time, these Unet approaches enhance forecast skill by approximately 5% to 25%, with performance naturally degrading from day 1 to day 3 and beyond. Temporal variability of the correlation shows that the Unet-log and Unet-GMM methods perform best, reaching correlation values above 0.95 at short lead times (6–24 hours), whereas the raw WRF rarely exceeds 0.93. This temporal variation may be partially influenced by more FZL values being at the surface during the 12Z forecasts, which results in closer agreement between model output and observations. Although the relative improvement is less pronounced during winter 2015-16, the overall improvement in FZL forecasts from the Unet models across all lead times remains evident (Figure 4). The correlation remains high,

File generated with AMS Word template 2.0

ranging from 0.87 to 0.97, while the RMSE reductions vary from 5% to 20%—demonstrating that, under both wet and dry winter conditions, Unet-based postprocessing models consistently improve forecast accuracy. At most of the lead times, Unet-Log and Unet-GMM (blue and red) outperform Unet-RMSE, reinforcing their superior skill in postprocessing WRF FZL forecasts.

In both winters, the raw WRF forecasts exhibit higher RMSE, ranging from 250 m to 460 m, particularly at longer lead times, when forecast uncertainty tends to increase. The FZL improvements by Unet models are consistent but more pronounced during the winter of 2016–17, where complex precipitation dynamics and enhanced moisture transport challenge dynamical models such as WRF. Conversely, in the 2015–16 winter season, the relative improvement from postprocessing is less prominent, suggesting that synoptic-scale variability and precipitation frequency influence the effectiveness of machine learning-based correction methods. The improvements in correlation and RMSE for Unet-GMM and Unet-Log, compared to raw WRF forecasts, remain robust at most lead times for both seasons, particularly within the first 72 hours. Beyond 72 hours, while improvements persist, the interquartile ranges of errors begin to overlap more frequently, reflecting greater uncertainty at longer lead times. The temporal variability of the correlation (Figures 3a and 4a) and RSME (Figures 3b and 4b) suggests that Unet-based methods perform best at most of the lead times. This improvement is consistent across both wet and dry years, underscoring the robustness of the Unet-based postprocessing in various atmospheric regimes. Notably, while all three Unet frameworks outperform the raw WRF forecasts, Unet-Log and Unet-GMM consistently yield the best results, highlighting the impact of different loss functions and distributional assumptions on postprocessing efficacy. Given this consistent trend, we will focus on the verification skill results from the Unet-GMM and Unet-Log in the remainder of the paper.
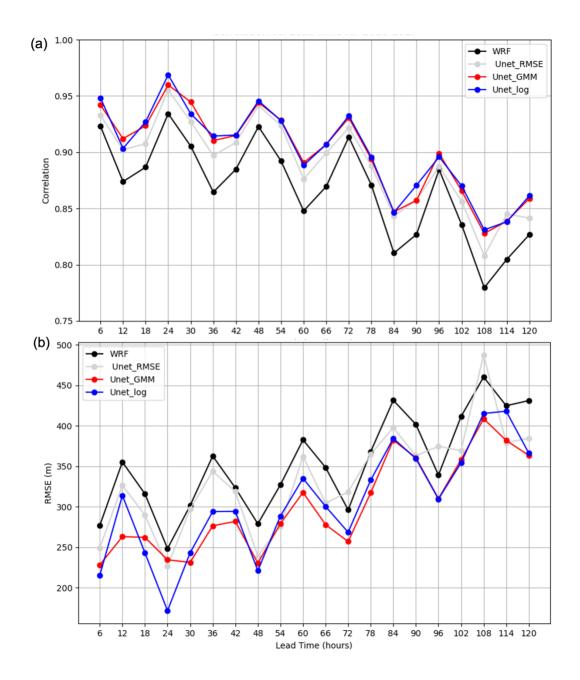
Figure 3. Performance metrics of 6 hourly WRF FZL and Unet postprocessing (Unet-RMSE, Unet-GMM, and Unet-log) forecasts during December 2016 - March 2017 over the Yuba-Feather Watershed: (a) Pearson correlation coefficient and (b) RMSE versus CNRFC observations as a function of lead time (hours).
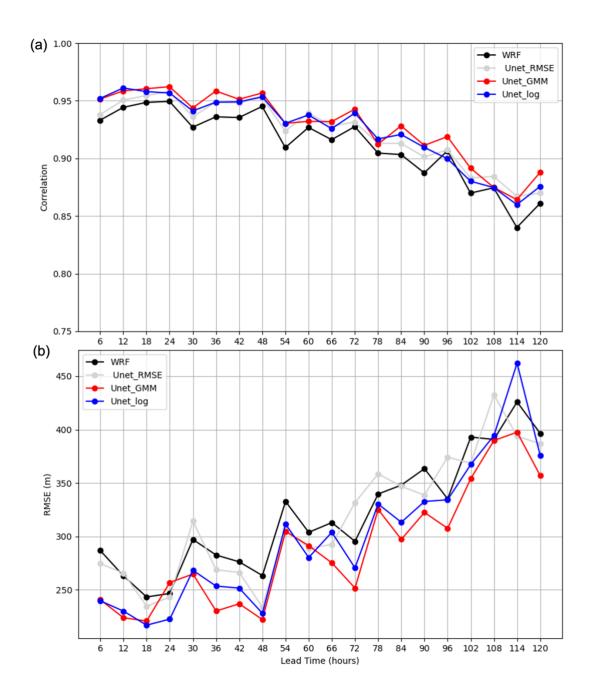
Figure 4. Same as figure 3, but for December 2015 - March 2016.

Figure 5 illustrates the spatial distribution of RMSE and correlation coefficients for Unet-GMM and Unet-log FZL forecasts, as well as the baseline WRF forecasts with respect to CNRFC observations for the winter of 2016-17 at a 2-day lead time. WRF exhibits high errors, particularly in the central and eastern regions of the Yuba-Feather watershed, with RMSE values ranging between 230 and 310 m (Figure 5a). These areas correspond to higher terrain, where FZLs are more frequently at the surface, potentially contributing to larger errors due to mismatches in vertical structure. However, both deep learning models, UNet-

14

GMM and UNet-log, significantly reduce RMSE over these regions to below 200 m in most of these areas (Figure 5b-c). Both Unet models also reduce the errors in the eastern parts of the watershed, a region characterized by complex terrain and high elevations (above 1,500 m), where freezing level forecasting remains challenging. Between the two postprocessing models, Unet-log reduces RMSE by an additional 10–15% compared to Unet-GMM for this lead time over most of the basin. The corresponding correlation maps (Figure 5d–f) display strong agreement between the FZL forecasts and the observed CNRFC freezing level, particularly in the western part of the watershed, where correlation values consistently exceed 0.85–0.90 across most grid points. Compared to WRF, which shows weaker correlations in certain areas in the eastern basin, both Unet models maintain a slightly higher spatial coherence with observations. Notably, the Unet postprocessing models achieve the lowest (highest) overall error (correlation), confirming that they not only enhance accuracy but also preserve spatial structure in FZL forecasts.

The results for the winter of 2015-16 exhibit elevated RMSE in the southwestern part of the basin, as shown in Figure 6. WRF forecasts show RMSE values exceeding 280 m in most areas, while Unet models reduce errors, particularly in the high-elevation eastern regions. Unet-GMM and Unet-log demonstrate notable improvements in FZL forecasts across the entire basin, especially in high altitudes (Figure 6b). This pattern aligns with the characteristics of a dry season, which has fewer precipitation events and generally higher freezing level elevations in the lower-elevation areas of the basin compared to a wetter season (2016-17 winter season). Correlation maps (Figure 6d–f) for the 2015-16 winter confirm that both deep learning models enhance spatial agreement with observations (reducing the RMSE) while maintaining correlation values consistently above 0.95 across most areas. These improvements are particularly evident in regions where WRF exhibits large biases, further underscoring the potential of machine learning approaches for improving numerical weather prediction model outputs. We have also examined other lead times and found that the improvement provided by the Unet postprocessing models remains consistent across most forecast horizons, further supporting their robustness in reducing errors. In the next section, we present the evaluation of a case study during an extreme AR event to demonstrate the effectiveness of the postprocessing models under extreme weather conditions.

As shown in Figures 5 and 6, the performance of the Unet models varies between seasons in different regions of the watershed. In the wetter 2016-17 winter, the greatest RMSE

reduction occurs in the lower-elevation western areas (~100–1,500 m), where Unet-log outperforms other models. However, in the drier winter of 2015-16, the Unet models primarily reduce errors in the higher-elevation eastern regions (above 1,500 m), suggesting a potential dependency of model performance on seasonal precipitation regimes and terrain complexity. In Section 4.3, we further evaluate the performance of the FZL forecasts by considering the model errors based on elevation. This will allow us to assess whether the improvements are consistent across different terrain levels or whether the model's effectiveness varies based on the underlying topography and meteorological conditions. By doing so, we aim to better understand the limitations and strengths of deep learning-based postprocessing in operational forecasting.
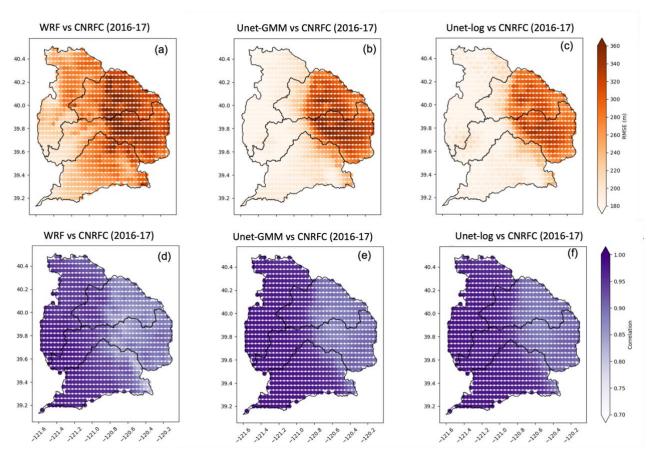

Figure 5. RMSE (a-c) and correlation (d-f) of 48-hour lead time FZL forecasts during winter 2016-2017 for (a and d) WRF, (b and e) UNet-GMM, and (c and f) UNet-log against observations from CNRFC.
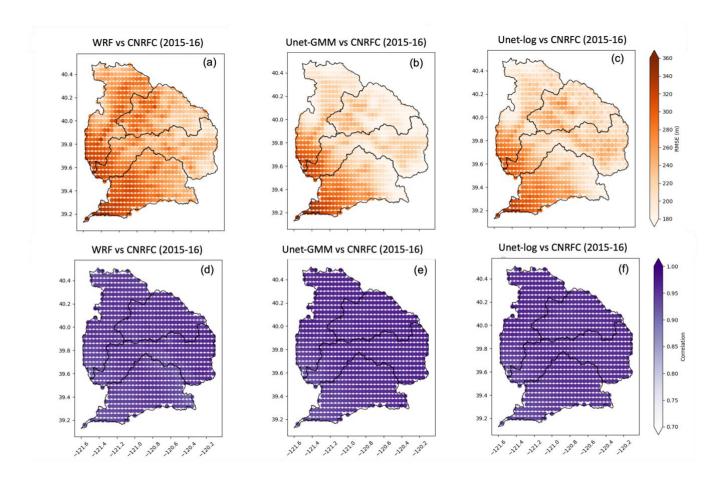
File generated with AMS Word template 2.0

Figure 6. Same as Figure 5 but for winter 2015-2016.

b. Case study: Atmospheric River event over the Western U.S., January 7 –9, 2017

During January 7–9, 2017, an AR5 event (based on the Ralph et al. 2019 AR Scale) impacted the Western U.S., producing heavy precipitation across the region. This event was characterized by an exceptionally strong moisture plume originating from the subtropical Pacific Ocean, which contributed to widespread hydrological impacts, including flooding and enhanced orographic precipitation in the Sierra Nevada. Integrated vapor transport (IVT) values exceeding 1000 kg m^{-1} s^{-1}, along with integrated water vapor (IWV) exceeding 30 mm made landfall over central California (Figure 7a-b). Accumulated precipitation during this event exceeded 200 mm over the Sierra Nevada and over 100 mm over the Coastal Ranges of Northern California (Figure 7c). Given the complexity of this event and its significant hydrometeorological impacts, it serves as an ideal case study to assess the performance of the Unet postprocessing models in improving FZL forecasts under these extreme weather conditions.
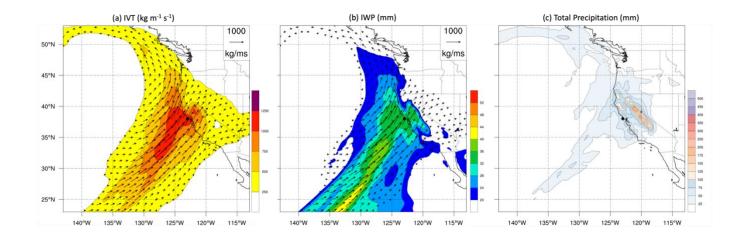
17

File generated with AMS Word template 2.0

Figure 7. ECMWF ERA5 IVT (kg m-1 s-1) with IVT vectors (a), IWV (mm) with IVT vectors (b) valid 15 UTC January 8, 2017, and total precipitation (mm) (c) for the AR5 event from 06 UTC January 7 – 08 UTC January 9, 2017.

To assess the performance of the postprocessing models during the AR5 event the spatial distribution of RMSE and correlation between the WRF, UNet-GMM, and UNet-log FZL forecasts and CNRFC observations for lead times of 24–48 hours (day 2) and 48–72 hours (day 3) were analyzed (Figure 8). The WRF model exhibits RMSE values exceeding 200 m for Day 2, particularly in the northwestern regions of the basin (Figure 8a). The Unet-GMM (Figure 8b) reduces errors across most of the study region with RMSE values dropping below 150 m, outperforming Unet-log (Figure 8c) for this lead time. For lead time Day 3 (Figure 8 d–f), RMSE overall increases across all models, as expected, due to forecast degradation at longer lead times. The WRF model exhibits RMSE values exceeding 550 m in most of the regions, particularly in the lower elevations (less than ~1500 m). The Unet-log and Unet-GMM models similarly lower RMSE values to below 430 m over this region. The Unet-GMM model exhibits slightly higher RMSE values than the Unet-log model but still improves upon the baseline WRF forecast. For Day 2, the spatial structure of correlation differs among the models. WRF shows a modest correlation, less than 0.73 in many parts of the basin, particularly in areas where RMSE is highest (Figure 9a). In contrast, Unet-GMM (Figure 9b) achieves relatively robust correlation values, exceeding 0.78 across most of the watershed, while Unet-log (Figure 9c) demonstrates improved agreement with observations over the southeast and northwest parts of the watershed when compared to WRF. The correlation maps do not show any specific spatial patterns, while forecast skill deteriorates with increasing lead time. For Day 3 (Figure 9d–f), Unet-log maintains relatively high correlations (~0.75–0.90) across much of the watershed, though performance in the southeast

18

and northwest is slightly reduced compared to Day 2. Unet-GMM also exhibits a different pattern from Day 2, with more uniform but slightly lower correlations, remaining around 0.75 across most areas.

Overall, the Unet postprocessing models reduce RMSE across all lead times, with error reductions exceeding 120 m in most areas compared to WRF while maintaining/improving correlation values, particularly at longer lead times. The spatial distribution of improvements varies during this AR5 event, with greater RMSE reductions in western high-elevation regions during Day 2, whereas low-elevation western areas show stronger improvements at longer lead times (Day 3) in comparison to WRF forecasts. These results reinforce the capability of Unet postprocessing to enhance deterministic weather model forecasts, particularly in the context of extreme atmospheric river events where accurate freezing level predictions are critical for hydrological forecasting and flood risk management. The next section evaluates model performance across different elevation bands to assess the consistency of improvements across varying topography.
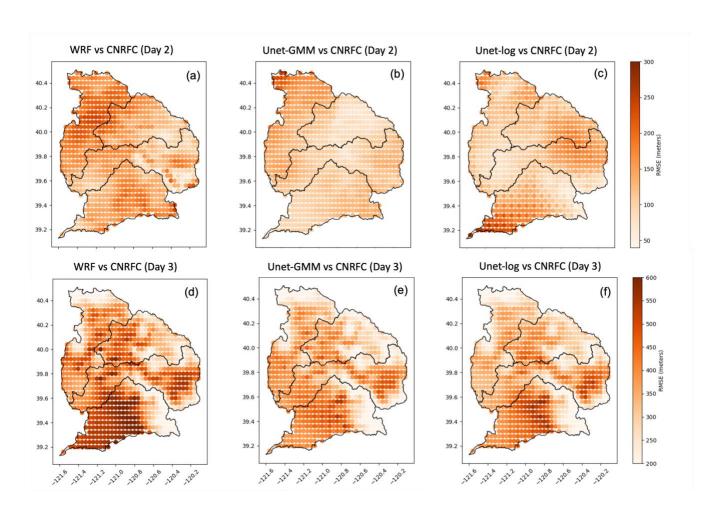
Figure 8. FZL forecasts RMSE for day 2 lead time (24 to 48 h) (a-c) and day 3 lead time (48 to 72 h) (d-f) for the case study (initialized on January 7, 2017) for (a and d) WRF, (b and e) UNet-GMM, and (c and f) UNet-log against observations from CNRFC.
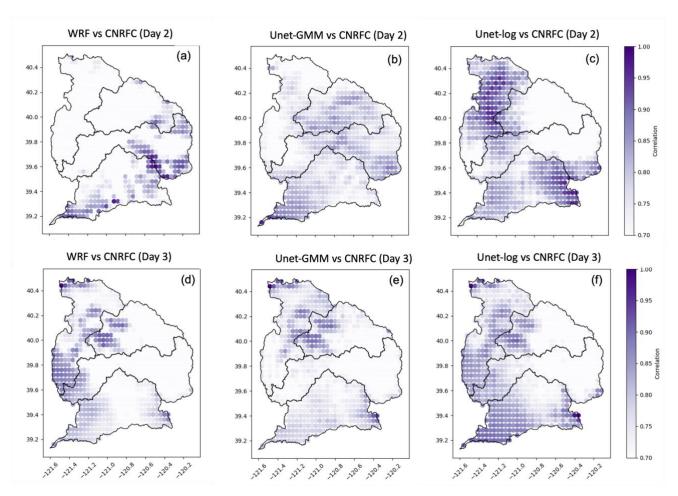

Figure 9. Same as Figure 8, but for the correlation coefficient.

c. *Performance of Unet postprocessed models across different elevational bands*

To evaluate the effectiveness of the Unet-based postprocessing models in varying topographic conditions, we assessed their performance across five elevation bands within the Yuba-Feather watershed. The following elevation bands were analyzed: 0–1000 m, 1000–1500 m, 1500–1750 m, 1750–2000 m, and >2000 m. These analyses were conducted for the winter seasons of 2016-17 and 2015-16. Grid points corresponding to each elevation band were selected, and the results were aggregated, with the number of grid points annotated in the corresponding plots. The panels are organized from the lowest elevation band at the bottom to the highest at the top to facilitate a direct comparison of model performance across terrain levels.

Across all elevation bands, a general decline in correlation is observed with increasing lead time, consistent with the expected loss of forecast skill. The Unet-GMM and Unet-Log models demonstrate improved correlation values relative to WRF across all lead times (6–120 hours), with the performance gap between WRF and the Unet-based models remaining consistent. At shorter lead times (6–48 hours), the Unet models achieve high correlations exceeding 0.80, comparable to and better than WRF. However, the performance gap becomes more pronounced beyond 48 hours. By 120 hours, the correlation for WRF drops below 0.70 in high-elevation bands, while Unet-based models maintain values above 0.75. Notably, in high-elevation regions (>1500 m), Unet models exhibit the most substantial improvements in skill, with correlation values ranging from 0.77 to 0.95, depending on lead time. In contrast, at lower elevations (<1500 m), the difference in RMSE and correlation between Unet and WRF is narrower, with Unet models achieving correlation values between 0.85 and 0.98. Across all elevations, the Unet models consistently reduce RMSE relative to WRF, with the most significant improvements observed in high-elevation bands and longer forecast horizons. At elevations above 1750 m, WRF's RMSE often exceeds 350 m beyond 72 hours, while Unet-GMM and Unet-Log reduce RMSE by approximately 50–150 m, with Unet-GMM achieving the lowest values throughout. This demonstrates the added value of the Unet postprocessing framework in capturing terrain-dependent forecast errors and improving skill across a range of conditions.

The results for the 2015–2016 winter season exhibit similar trends, albeit with a smaller performance gap between the raw WRF forecasts and the postprocessed outputs from the Unet models (Figure 11). While both Unet models still enhance forecast skill, their advantage over WRF is slightly less pronounced than in the following winter (2016–2017), particularly in the mid and high-elevation bands (>1500 m). Interestingly, for the 2015–2016 season, the correlation values remain above 0.8, while the RMSE values do not fluctuate significantly across different elevation bands. This finding suggests that the performance of the two postprocessing models is influenced by year-to-year variability, synoptic conditions, and seasonal characteristics. However, across both seasons, the Unet postprocessing models consistently improve correlation and reduce RMSE compared to WRF, demonstrating their robustness in enhancing FZL forecasts across diverse topographic conditions and across different winter seasons.

File generated with AMS Word template 2.0
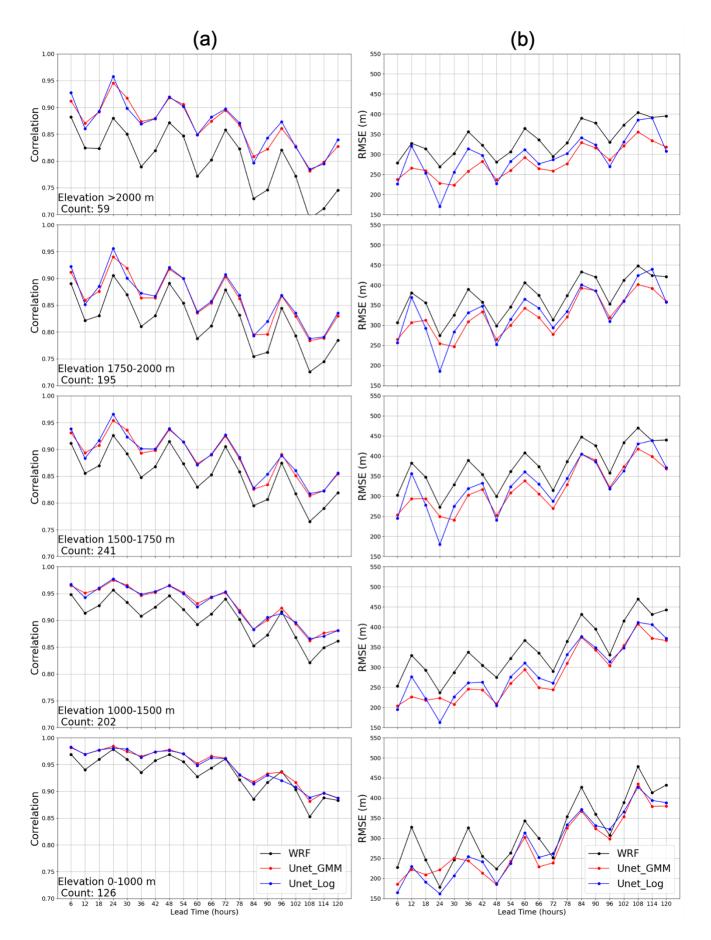
File generated with AMS Word template 2.0

Figure 10. Performance of 6 hourly WRF FZL and two Unet postprocessing forecasts over different elevation ranges during December 2016 - March 2017: (a) correlation coefficient and (b) RMSE versus CNRFC observations as a function of lead time (hours).

File generated with AMS Word template 2.0

Figure 11. Same as Figure 5 but for winter 2015-2016.

5. Summary and Conclusions

This study demonstrates the potential of applying deep learning-based postprocessing, using a Unet architecture, to refine freezing level (FZL) forecasts derived from West-WRF. Based on seven winter seasons of 6-hourly CNRFC FZL observations for training and two winters (2015–2016 and 2016–2017) for detailed evaluation, the Unet models enhance West-WRF outputs in forecasting FZL variations across the Yuba-Feather River watershed. This enhancement persists across different topographic regions, in an AR case study, and at forecast lead times up to 5 days, underscoring the value of advanced machine learning frameworks for weather prediction in complex terrain. This postprocessing method can be extended to other watersheds in the Sierra Nevada and beyond. The methodology utilizes the Unet architecture to leverage spatial features in meteorological fields while using Log-Cosh Error and Gaussian Mixture Model (GMM) loss functions, allowing the postprocessing framework to capture a broader range of error distributions than a conventional RMSE-based approach.

Results over the wet winter of 2016–2017 show substantial reductions in RMSE (15–25%) and increases in correlation relative to West-WRF, particularly in higher-elevation areas and beyond 72-hour lead times when raw forecast error grows. Although the drier winter of 2015–2016 exhibits less dramatic overall improvements, Unet-GMM and Unet-Log still consistently improve upon the raw West-WRF, highlighting the robustness of deep learning postprocessing in different precipitation regimes. A case study of a high-impact atmospheric river (AR5; Ralph et al. 2019) event from January 7–9, 2017, demonstrates that Unet-based models effectively capture complex FZL fluctuations and reduce forecast errors, even under extreme moisture transport and intense precipitation. Evaluations across five elevational bands reveal that deep learning corrections are especially beneficial at altitudes above 1500 m, where weather variability, precipitation phase transitions, and terrain-induced processes pose additional challenges for deterministic numerical weather models. Error reductions are also evident at lower elevations and extended lead times, aligning with the broader finding that incorporating ML methods can systematically address biases in physically based models. Compared to uncorrected West-WRF forecasts, the Unet-based models reduce RMSE, deliver higher correlation coefficients, and more accurately capture

25

the elevations corresponding to the rain–snow transition zone (approximately 1000–2500 m), in alignment with recent studies (Bair et al. 2021; Yang et al. 2022).

From an operational standpoint, improving FZL forecasts is critical for water resource management and optimal hydrological forecasting. FIRO efforts, particularly in AR-prone watersheds, rely on accurate precipitation phase predictions to optimize flood control and water storage. Encouraged by these findings, future research will examine how these FZL forecast enhancements propagate to inflow forecasts and, ultimately, reservoir decision-making. Following approaches similar to Sumargo et al. (2020), a quantitative assessment of how FZL error reductions affect reservoir operations will be undertaken—an essential step toward integrating deep learning postprocessing into hydrologic and operational models. Looking ahead, several avenues exist for enhancing the efficacy and robustness of Unet-based postprocessing. These include expanding to probabilistic frameworks for capturing forecast uncertainties (such as the study by Hu et al., 2023 for precipitation postprocessing), exploring training Unet frameworks for individual watersheds or transfer learning to adapt models to other mountain basins, and incorporating additional atmospheric variables (e.g., humidity profiles, wind speeds) into the input feature space. Such extensions are poised to further boost predictive skill and resilience under a variety of weather conditions.

Overall, this study offers a promising advancement for applying modern deep learning architectures to resolve long-standing challenges in complex-terrain FZL forecasting. By systematically reducing model errors, enhancing spatial coherence, and improving lead-time performance, Unet-based postprocessing demonstrates the potential to pave the way for more accurate and adaptive hydrometeorological services, particularly important in a changing climate where extreme precipitation events are growing in frequency and intensity (Espinoza et al. 2018).

Data Availability Statement.

The California Nevada River Forecast Center (CNRFC) gridded freezing level observations are publicly available at

File generated with AMS Word template 2.0

APPENDIX A

Comparison of Accumulated Winter Precipitation from ERA5 Reanalysis

This appendix presents accumulated precipitation from the European Centre for Medium-Range Weather Forecasts' fifth-generation reanalysis product (ERA)to illustrate the contrast between the wet and dry winters used in the model test period: December 2015–March 2016 and December 2016–March 2017. These figures support the discussion in the last paragraph of Section 3.3. Figure A1a shows the total accumulated precipitation for the winter of 2015–2016, which was characterized by below-average precipitation despite a strong El Niño event. Figure A1b displays the accumulated precipitation for the winter of 2016–2017, a season noted for record-breaking rainfall across California. To better highlight spatial differences between the two winters, Figure A1c shows the difference in accumulated precipitation (2016–17 minus 2015–16). These figures provide additional context for interpreting model performance under contrasting hydroclimatic conditions.
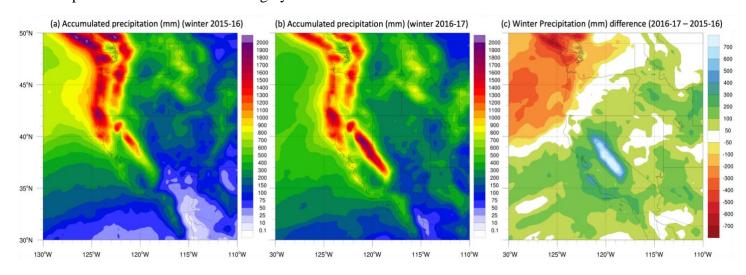

Figure A1. ERA5 accumulated precipitation (mm): (a) from December 1, 2015, to March 31, 2016; (b) from December 1, 2016, to March 31, 2017; (c) difference in accumulated precipitation (mm) between the two test winters (2016–2017 minus 2015–2016).

APPENDIX B

File generated with AMS Word template 2.0

Verification metrics

The performance of deterministic postprocessing models is evaluated with Pearson correlation coefficient (higher the better, perfect value is 1) and root-mean-square error (RMSE, lower the better, ideal value is 0):

$$Correlation = \frac{1}{n} \frac{\sum_{i=1}^{n}(P_i - \underline{P})(O_i - \underline{O})}{\sqrt{\sum_{i=1}^{n}(P_i - \underline{P})^2}\sqrt{\sum_{i=1}^{n}(O_i - \underline{O})^2}}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2}$$

Where P_i represents the ith FZL forecasts, O_i refers to the corresponding FZL observation, with n denoting the total number of samples.

REFERENCES

Alerskans, E., J. Nyborg, M. Birk, and E. Kaas, 2022: A transformer neural network for predicting near-surface temperature. *Meteorological Applications*, **29**, https://doi.org/10.1002/met.2098.

Badrinath, A., L. D. Monache, N. Hayatbini, W. Chapman, F. Cannon, and M. Ralph, 2023: Improving Precipitation Forecasts with Convolutional Neural Networks. *Weather Forecast*, **38**, 291–306, https://doi.org/10.1175/WAF-D-22-0002.1.

Bair, E. H., T. Stillinger, and J. Dozier, 2021: Snow Property Inversion From Remote Sensing (SPIReS): A Generalized Multispectral Unmixing Approach With Examples From MODIS and Landsat 8 OLI. *IEEE Transactions on Geoscience and Remote Sensing*, **59**, 7270–7284, https://doi.org/10.1109/TGRS.2020.3040328.

Cannon, F., F. M. Ralph, A. M. Wilson, and D. P. Lettenmaier, 2017: GPM satellite radar measurements of precipitation and freezing level in atmospheric rivers: Comparison with ground-based radars and reanalyses. *Journal of Geophysical Research: Atmospheres*, **122**, 12,747-12,764, https://doi.org/10.1002/2017JD027355.

Chapman, W. E., A. C. Subramanian, L. Delle Monache, S. P. Xie, and F. M. Ralph, 2019: Improving Atmospheric River Forecasts With Machine Learning. *Geophys Res Lett*, **46**, 10627–10635, https://doi.org/10.1029/2019GL083662.

Chapman, W. E., L. D. Monache, S. Alessandrini, A. C. Subramanian, F. Martin Ralph, S. P. Xie, S. Lerch, and N. Hayatbini, 2022: Probabilistic Predictions from Deterministic

File generated with AMS Word template 2.0

Atmospheric River Forecasts with Deep Learning. *Mon Weather Rev*, **150**, 215–234, https://doi.org/10.1175/MWR-D-21-0106.1.

Cho, D., C. Yoo, B. Son, J. Im, D. Yoon, and D. H. Cha, 2022: 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 Clim Extrem*, **35**, 100410, https://doi.org/10.1016/J.WACE.2022.100410.

Cobb, A., and Coauthors, 2023: West-WRF 34-Year Reforecast: Description and Validation. *J Hydrometeorol*, **24**, 2125–2140, https://doi.org/10.1175/JHM-D-22-0235.1.

Delle Monache, L., T. Nipen, Y. Liu, G. Roux, and R. Stull, 2011: Kalman Filter and Analog Schemes to Postprocess Numerical Weather Predictions. *Mon Weather Rev*, **139**, 3554–3570, https://doi.org/10.1175/2011MWR3653.1.

——, F. Anthony Eckel, D. L. Rife, B. Nagarajan, and K. Searight, 2013: Probabilistic Weather Prediction with an Analog Ensemble. *Mon Weather Rev*, **141**, 3498–3516, https://doi.org/10.1175/MWR-D-12-00281.1.

——, and Coauthors, 2024: CW3E's West-WRF 200-member Ensemble. *Mon Weather Rev*.

Dettinger, M., 2011: Climate Change, Atmospheric Rivers, and Floods in California – A Multimodel Analysis of Storm Frequency and Magnitude Changes1. *JAWRA Journal of the American Water Resources Association*, **47**, 514–523, https://doi.org/10.1111/J.1752-1688.2011.00546.X.

——, and D. R. Cayan, 2014: Drought and the California delta - A matter of extremes. *San Francisco Estuary and Watershed Science*, **12**, 1–6, https://doi.org/10.15447/SFEWS.2014V12ISS2ART4.

Espinoza, V., D. E. Waliser, B. Guan, D. A. Lavers, and F. M. Ralph, 2018: Global Analysis of Climate Change Projection Effects on Atmospheric Rivers. *Geophys Res Lett*, **45**, 4299–4308, https://doi.org/10.1029/2017GL076968.

Ghazvinian, M., Y. Zhang, D. J. Seo, M. He, and N. Fernando, 2021: A novel hybrid artificial neural network - Parametric scheme for postprocessing medium-range precipitation forecasts. *Adv Water Resour*, **151**, https://doi.org/10.1016/j.advwatres.2021.103907.

——, ——, T. M. Hamill, D. J. Seo, and N. Fernando, 2022: Improving Probabilistic Quantitative Precipitation Forecasts Using Short Training Data through Artificial Neural Networks. *J Hydrometeorol*, **23**, 1365–1382, https://doi.org/10.1175/JHM-D-22-0021.1.

——, and Coauthors, 2024: Deep Learning of a 200-Member Ensemble with a Limited Historical Training to Improve the Prediction of Extreme Precipitation Events. *Mon Weather Rev*, **152**, 1587–1605, https://doi.org/10.1175/MWR-D-23-0277.1.

Glahn, H. R., and D. A. Lowry, 1962: *The Use of Model Output Statistics (MOS) in Objective Weather Forecasting*. 1203–1211pp.

Henn, B., R. Weihs, A. C. Martin, F. Martin Ralph, and T. Osborne, 2020: Skill of rain-snow level forecasts for landfalling atmospheric rivers: A multimodel assessment using california's network of vertically profiling radars. *J Hydrometeorol*, **21**, 751–771, https://doi.org/10.1175/JHM-D-18-0212.1.

Hu, W., M. Ghazvinian, W. E. Chapman, A. Sengupta, F. M. Ralph, and L. D. MonachE, 2023: Deep Learning Forecast Uncertainty for Precipitation over the Western United States. *Mon Weather Rev*, **151**, 1367–1385, https://doi.org/10.1175/MWR-D-22-0268.1.

Jadon, A., A. Patil, and S. Jadon, 2024: A Comprehensive Survey of Regression-Based Loss Functions for Time Series Forecasting. 117–147, https://doi.org/10.1007/978-981-97-3245-6_9/TABLES/20.

Jasperse, J., and Coauthors, 2020: *Lake Mendocino forecast informed reservoir operations final viability assessment*. 131pp., accessed 27 June 2024, https://escholarship.org/uc/item/3b63q04n.

Jasperse, J. and R. F. and A. M. and B. L. and D. M. and D. M. and H. R. and J. C. and R. P. and T. C. and others, 2015: *A comprehensive plan to evaluate the viability of forecast informed reservoir operations for Lake Mendocino*. 1–374pp.

Kumar, A., and M. Chen, 2017: What is the variability in US west coast winter precipitation during strong El Niño events? *Clim Dyn*, **49**, 2789–2802, https://doi.org/10.1007/S00382-016-3485-9/FIGURES/9.

Li, W., B. Pan, J. Xia, and Q. Duan, 2022: Convolutional neural network-based statistical post-processing of ensemble precipitation forecasts. *J Hydrol (Amst)*, **605**, https://doi.org/10.1016/j.jhydrol.2021.127301.

Luna-Niño, R., A. Gershunov, F. M. Ralph, A. Weyant, K. Guirguis, M. J. DeFlorio, D. R. Cayan, and A. P. Williams, 2025: Heresy in ENSO teleconnections: atmospheric rivers as disruptors of canonical seasonal precipitation anomalies in the Southwestern US. *Clim Dyn*, **63**, 1–11, https://doi.org/10.1007/S00382-025-07583-1/FIGURES/5.

Minder, J. R., and D. E. Kingsmill, 2013: Mesoscale variations of the atmospheric snow line over the northern sierra nevada: Multiyear statistics, case study, and mechanisms. *J Atmos Sci*, **70**, 916–938, https://doi.org/10.1175/JAS-D-12-0194.1.

Neiman, P. J., F. M. Ralph, G. A. Wick, J. D. Lundquist, and M. D. Dettinger, 2008: Meteorological characteristics and overland precipitation impacts of atmospheric rivers affecting the West coast of North America based on eight years of SSM/I satellite observations. *J Hydrometeorol*, **9**, 22–47, https://doi.org/10.1175/2007JHM855.1.

Palerme, C., and Coauthors, 2024: Improving short-term sea ice concentration forecasts using deep learning. *Cryosphere*, **18**, 2161–2176, https://doi.org/10.5194/tc-18-2161-2024.

Pan, B., K. Hsu, A. AghaKouchak, and S. Sorooshian, 2019: Improving Precipitation Estimation Using Convolutional Neural Network. *Water Resour Res*, **55**, 2301–2321, https://doi.org/10.1029/2018WR024090.

Qin, Y., B. Li, X. Sun, Y. Chen, and X. Shi, 2019: Nonlinear response of runoff to atmospheric freezing level height variation based on hybrid prediction models. *Hydrological Sciences Journal*, **64**, 1556–1572, https://doi.org/10.1080/02626667.2019.1662023.

Ralph, F. M., P. J. Neiman, G. A. Wick, F. M. Ralph, P. J. Neiman, and G. A. Wick, 2004: Satellite and CALJET Aircraft Observations of Atmospheric Rivers over the Eastern North Pacific Ocean during the Winter of 1997/98. *MWRv*, **132**, 1721, https://doi.org/10.1175/1520-0493(2004)132.

——, ——, and R. Rotunno, 2005: Dropsonde Observations in Low-Level Jets over the Northeastern Pacific Ocean from CALJET-1998 and PACJET-2001: Mean Vertical-Profile and Atmospheric-River Characteristics. *Mon Weather Rev*, **133**, 889–910, https://doi.org/10.1175/MWR2896.1.

——, J. J. Rutz, J. M. Cordeira, M. Dettinger, M. Anderson, D. Reynolds, L. J. Schick, and C. Smallcomb, 2019: A Scale to Characterize the Strength and Impacts of Atmospheric

Rivers. *Bull Am Meteorol Soc*, **100**, 269–289, https://doi.org/10.1175/BAMS-D-18-0023.1.

Ronneberger, O., P. Fischer, and T. Brox, 2015: U-Net: Convolutional Networks for Biomedical Image Segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **9351**, 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.

Siler, N., Y. Kosaka, S. P. Xie, and X. Li, 2017: Tropical ocean contributions to California's surprisingly dry El Niño of 2015/16. *J Clim*, **30**, 10067–10079, https://doi.org/10.1175/JCLI-D-17-0177.1.

Sumargo, E., F. Cannon, F. M. Ralph, and B. Henn, 2020: Freezing Level Forecast Error Can Consume Reservoir Flood Control Storage: Potentials for Lake Oroville and New Bullards Bar Reservoirs in California. *Water Resour Res*, **56**, https://doi.org/10.1029/2020WR027072.

Sun, L., Y. Lan, X. Sun, X. Liang, J. Wang, Y. Su, Y. He, and D. Xia, 2024: Deterministic Forecasting and Probabilistic Post-Processing of Short-Term Wind Speed Using Statistical Methods. *Journal of Geophysical Research: Atmospheres*, **129**, https://doi.org/10.1029/2023JD040134.

Taillardat, M., O. Mestre, M. Zamo, and P. Naveau, 2016: Calibrated Ensemble Forecasts Using Quantile Regression Forests and Ensemble Model Output Statistics. *Mon Weather Rev*, **144**, 2375–2393, https://doi.org/10.1175/MWR-D-15-0260.1.

Talbot, C. A., Ralph F. M., and Jasperse J., 2019: Forecast-informed reservoir operations: Lessons learned from a multi-agency joint research and operations effort. *Federal Interagency Sedimentation and Hydrologic Modeling Conf., Reno, NV, SEDHYD*.

Vannitsem, S., and Coauthors, 2021: Statistical postprocessing for weather forecasts review, challenges, and avenues in a big data world. *Bull Am Meteorol Soc*, **102**, E681–E699, https://doi.org/10.1175/BAMS-D-19-0308.1.

Wilks, D. S., and T. M. Hamill, 2007: Comparison of Ensemble-MOS Methods Using GFS Reforecasts. *Mon Weather Rev*, **135**, 2379–2390, https://doi.org/10.1175/MWR3402.1.

Yang, K., K. N. Musselman, K. Rittger, S. A. Margulis, T. H. Painter, and N. P. Molotch, 2022: Combining ground-based and remotely sensed snow data in a linear regression

model for real-time estimation of snow water equivalent. *Adv Water Resour*, **160**, 104075, https://doi.org/10.1016/J.ADVWATRES.2021.104075.

Zhang, T., Z. Liang, W. Li, J. Wang, Y. Hu, and B. Li, 2023: Statistical post-processing of precipitation forecasts using circulation classifications and spatiotemporal deep neural networks. *Hydrol Earth Syst Sci*, **27**, 1945–1960, https://doi.org/10.5194/hess-27-1945-2023.

Zhu, Y., and R. E. Newell, 1994: Atmospheric rivers and bombs. *Geophys Res Lett*, **21**, 1999–2002, https://doi.org/10.1029/94GL01710.

——, and R. E. Newell, 1998: A Proposed Algorithm for Moisture Fluxes from Atmospheric Rivers. *Monthly Weather Review* , **126**, 725–735.