Geophysical Research Letters

RESEARCH LETTER
10.1029/2019GL083662

Improving Atmospheric River Forecasts With Machine Learning

W. E. Chapman[1], A. C. Subramanian[2], L. Delle Monache[1], S. P. Xie[1], and F. M. Ralph[1]

[1]Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA, USA, [2]Atmospheric and Oceanic Sciences, University of Colorado Boulder, Boulder, CO, USA

Key Points:
- The GFS forecast field of integrated vapor transport is used for a convolutional neural network-based forecast postprocessing method
- The machine learning algorithm reduces the full-field root-mean-square error and improves the correlation with ground truth
- An error deconstruction shows that the dominant improvements come from the reduction of random error and conditional biases

Supporting Information:
- Supporting Information S1

Correspondence to:
W. E. Chapman,
wchapman@ucsd.edu

Citation:
Chapman, W. E., Subramanian, A. C., Delle Monache, L., Xie, S. P., & Ralph, F. M. (2019). Improving atmospheric river forecasts with machine learning. *Geophysical Research Letters*, *46*□ □□□□□□□□□□□https://doi.org/10.1029/2019GL083662

Received 15 MAY 2019
Accepted 21 AUG 2019
Accepted article online 26 AUG 2019
/ ĘčċąĖĄąĀĒďċąďā □2$/ □□□□

Abstract This study tests the utility of convolutional neural networks as a postprocessing framework for improving the National Center for Environmental Prediction's Global Forecast System's integrated vapor transport forecast field in the Eastern Pacific and western United States. Integrated vapor transport is the characteristic field of atmospheric rivers, which provide over 65% of yearly precipitation at some western U.S. locations. The method reduces full-field root-mean-square error (RMSE) at forecast leads from 3 hr to seven days (9–17% reduction), while increasing correlation between observations and predictions (0.5–12% increase). This represents an approximately one- to two-day lead time improvement in RMSE. Decomposing RMSE shows that random error and conditional biases are predominantly reduced. Systematic error is reduced up to five-day forecast lead, but accounts for a smaller portion of RMSE. This work demonstrates convolutional neural networks potential to improve forecast skill out to seven days for precipitation events affecting the western United States.

Plain Language Summary Machine learning methods are data-driven algorithms that improve by examining massive amounts of existing data. We explore the utility of a computer-vision machine learning technique to reduce error in numerical weather forecasts of the characteristic field for atmospheric rivers (ARs). ARs are long narrow corridors of anomalous vapor transport capable of providing both beneficial and hazardous precipitation. Therefore, accurately forecasting AR events is extremely important from a water supply and flood protection standpoint. We show significant forecast improvements by applying machine learning postprocessing for lead times ranging from 3 hr to seven days, making the predictions more valuable to stakeholders affected by AR events.

1. Introduction

Numerical weather prediction (NWP) models provide the atmospheric variables necessary to determine projected atmospheric states, based on a numerical integration of a discretized version of the Navier-Stokes equations (Richardson, 1922). However, due to uncertainty in initial conditions, numerical approximation, and model deficiencies, error increases nonlinearly and NWP forecast skill decreases with model time integration (Lorenz, 1963). Statistical forecast postprocessing techniques, which utilize historical forecasts and observations to correct for error in current predictions, have been found to significantly improve forecast skill across multiple atmospheric variables. Algorithms developed to determine and correct for NWP error include model output statistics approaches (e.g., Carter et al., 1989; Glahn & Lowry, 1972; Wilks & Hamill, 2007), running mean techniques (e.g., Hacker & Rife, 2008; Stensrud & Skindlov, 2002; Stensrud & Yussouf, 2003), algorithms based on Kalman filtering (e.g., Delle Monache et al., 2006; Homleid, 1995; McCollor & Stull, 2008; Roeger et al., 2003), and analog-based methods which draw from past events to match designed features of the current forecast to correct it (Delle Monache et al., 2011).

The North American West Coast presents a challenge in water forecasting. Wintertime precipitation provides almost all the annual input to the water budget, generally within a few large horizontal vapor transport events (Dettinger et al., 2011) termed atmospheric rivers (ARs). ARs are long (>2,000 km) and narrow (<1,000 km) corridors of anomalous vapor transport, typically associated with a low-level jet, ahead of the cold section of an extratropical cyclone (e.g., Dacre et al., 2015; Sodemann & Stohl, 2013; Warner et al., 2012), which deliver the majority of poleward vapor transport (>90%) in less than 10% of the zonal circumference of the extratropics (Ralph et al., 2004; Zhu & Newell, 1998). Vertically integrated vapor transport (IVT) is the characteristic metric which defines the strength of an AR (Ralph et al., 2018). IVT is a

combined thermodynamic and momentum metric which integrates specific humidity and zonal and meridional components of the wind from 1,000 to 300 hPa.

ARs contribute 30–65% of annual precipitation on the U.S. West Coast, and ARs contribute 60–100% of the most extreme North American West Coast hydrometeorological events (Gershunov et al., 2017; Lamjiri et al., 2017). Lavers et al. (2016) found that IVT evolution is dominated by synoptic-scale processes, and thus has a higher predictability than precipitation, which depends more on mesoscale and microphysical processes. Therefore, at long lead times, forecasting IVT, rather than precipitation, may be more valuable to water management and hazard mitigation. However, forecasting for AR events has proved difficult. A study by Wick et al. (2013) examined the National Centers of Environmental Prediction's Global Forecast System (GFS) West Coast forecast skill over the Northeast Pacific across three cold seasons and found that average AR landfall location errors were approximately 600 km at seven-day lead time.

We propose a novel NWP postprocessing technique, applied to the IVT field, that leverages a subclass of machine learning computer vision techniques: convolutional neural networks (CNN). CNNs are able to encode features from an input field, at varying spatial scales and levels of abstraction (Bengio, 2009; Hinton et al., 2006), which maximize predictive skill to a specified output field. These networks are adept at processing large and complex data sets and determining meaningful relationships. CNNs have proven to be extremely successful at image recognition, semantic segmentation, image denoising, and image super resolution (Bojarski et al., 2017; Dong et al., 2014; He et al., 2015; Long et al., 2015; Zhang et al., 2017). CNNs are well suited to atmospheric fields, where systems across multiple scales govern atmospheric flow.

More recently, flexible forecast prediction and postprocessing approaches based on artificial neural networks, which take advantage of increased computational power to learn from a large database of past forecasts, have been proposed (e.g., Tao et al., 2016). Neural networks reduced bias and improved ensemble 2-m temperature prediction over Germany (Rasp & Lerch, 2018). Random forests have been used for storm-based probabilistic hail forecasting (e.g., Gagne et al., 2017). When combined with the physical understanding of atmospheric processes, machine learning has been shown to aid in high-impact weather decision making (McGovern et al., 2017). Specifically, CNNs are beginning to be used for scientific discovery and forecasting and have emerged as diagnostic tools for determining important atmospheric variables across scales (e.g., Kurth et al., 2018; Toms et al., 2019). CNNs have been utilized to provide forecast uncertainty estimates upon initialization (Sebastian Scher & Messori, 2018). Additionally, purely CNN-based forecast methods have arisen for prediction and nowcasting applications, relying on data alone to mimic atmospheric dynamics (Shi et al., 2015; Dueben & Bauer, 2018; S. Scher, 2018; Sebastian Scher & Messori, 2019). This study aims to extend the utility of CNNs as a postprocessing method to improve predictions up to seven days ahead.

At every forecast lead time, we create a new CNN which inputs a GFS IVT magnitude forecast field and outputs a corrected IVT forecast field. The present study evaluates whether historical forecast error can be used in conjunction with CNNs as a postprocessing tool to improve short- and medium-range IVT forecasts.

2. Data and Methodology

2.1. Forecasts

GFS predictions (Moorthi et al., 2001) at a 0.5-degree horizontal spatial resolution on 64 vertical levels for daily 0000 and 1200 UTC model initializations are utilized to calculate forecasted IVT. Here the forecasts from 3 to 168 hr are examined (3-hr increments for the first 12-hr period, 12-hr increment for the following day, and 24-hr increments for the remaining 168-hr forecast lead times) for the cold season (defined here as October–April) from 2006 to 2018. This includes ~5,000 data fields for every forecast lead time or ~55,000 forecasted fields across all lead times. This study's region of interest spans coastal North America and the Eastern Pacific from 180°W to 110°W longitude, and 10°N to 60°N latitude.

2.2. Ground Truth

IVT from the National Aeronautics and Space Administration's Modern-Era Retrospective Analysis for Research and Applications version 2 (MERRA-2) reanalysis is used as ground truth to diagnose forecast error and CNN model training. MERRA-2 provides a regularly gridded record of the global atmosphere, including assimilated satellite, surface station, wind profiler, radio occultation, and radiosonde observations. MERRA-2 data are resolved on a 0.625×0.5-degree grid and interpolated to 21-pressure levels between 1,000 and 300

hPa for IVT calculations (Gelaro et al., 2017; McCarty et al., 2016). For consistency, GFS IVT is regridded and upscaled to MERRA-2 resolution using a first- and second-order conservative remapping scheme (Schulzweida, 2019).

2.3. Methodology and Experimental Design

We compare four separate forecasts to examine the relative skill of the CNN postprocessing: (1) GFS is used as the dynamical NWP model and provides a deterministic forecast of future IVT states from current meteorological observations; (2) a climatological forecast created from a 21-day running mean, centered on the forecast day of interest, from MERRA-2 IVT fields spanning 1980–2018; (3) a persistence forecast created by repeating the GFS analysis at 0-hr lead for every lead time; and (4) a forecast derived from postprocessing the GFS IVT forecast with a CNN (hereafter referred to as ARcnn when referencing the architecture and ARcnn-IVT for the forecast).

2.4. Convolutional Neural Networks and the Network Used in This Study

Neural networks are known to be able to approximate nonlinear functions (Nielsen, 2015). CNNs are a class of neural network, in which multiple layers of optimized functions map input data fields (GFS forecasts in this study) to an output (ARcnn-IVT). CNNs use convolutional kernels to propagate images from one layer to the next. Each convolutional kernel is trained to highlight important image features. Following each convolutional layer, nonlinearities are introduced, which operate on every produced feature map. The ARcnn architecture was inspired by a class of CNNs termed *denoising autoencoders* (Vincent et al., 2008).

Denoising autoencoders are trained, with coupled pairs of noisy and clean images, taking a noise corrupted image and removing that noise. Here GFS IVT forecasts are treated as noisy images, the noise representing the prediction error, and ARcnn corrects the forecast toward a clean image, MERRA-2 ground truth. ARcnn contains no compression or pooling information layers which reduce dimensionality. Therefore, a consistent dimension (determined by the latitude and longitude points of the region of interest) is retained throughout the network and in the prediction.

The optimization of the kernel filter weights occurs iteratively, in which each iteration finds the weights of the functions to minimize the loss between the output (ARcnn-IVT) and a desired field (MERRA-2). ARcnn utilizes an Adam optimizer (Kingma et al., 2014) with a learning rate that decreased from 0.001 to $5e^{-6}$ upon validation plateaus and batch size 20. The error is determined between the network forecast and the ground truth data, and the gradient of the error field is calculated for each kernel weight of the network. The model weights update each iteration by stepping in the direction opposite of this gradient. ARcnn optimized utilizing mean-squared error loss. Once trained, ARcnn produces an estimated IVT field that has learned error from previous forecasts and has the ability to correct a portion of these errors. A detailed description of CNNs and the ARcnn model architecture is in the supporting information. For further information on CNNs the reader is referred to Nielsen (2015).

GFS forecasts were separated by date into training (October 2008 to April 2016), validation (October 2016 to April 2017), and testing (October 2017 to April 2018) data sets. Training data are shown iteratively to the neural network to optimize CNN model weights. Validation data are used to compute performance metrics during training. Testing data are unseen by the network and utilized only for evaluating the postprocessing skill. The final year of data (October 2017 to April 2018) is reserved for testing and is independent from any training data. Each lead time in the testing period consists of ~450 forecasts. Table S1 shows the number of samples and the frequency of ARs in the training, validation, and testing data sets. Each forecast lead is trained, validated, and tested on ~5,000 forecasts. A new CNN is created and trained for each forecast lead time. However, across these CNNs, there is valuable similarity in the IVT feature detection during convolution. To exploit this similarity during training, we utilized a sequential training scheme in which the model network weights from previous forecast lead times initialized network weights at subsequent forecast leads. This decreased the number of model training cycles and improved total error testing results (not shown).

Table S2 summarizes the model architecture and training parameters. An exhaustive number of training cycles, using common CNN model parameters, was performed to determine optimal model settings. The final parameters were selected by choosing the configuration with the lowest validation error.

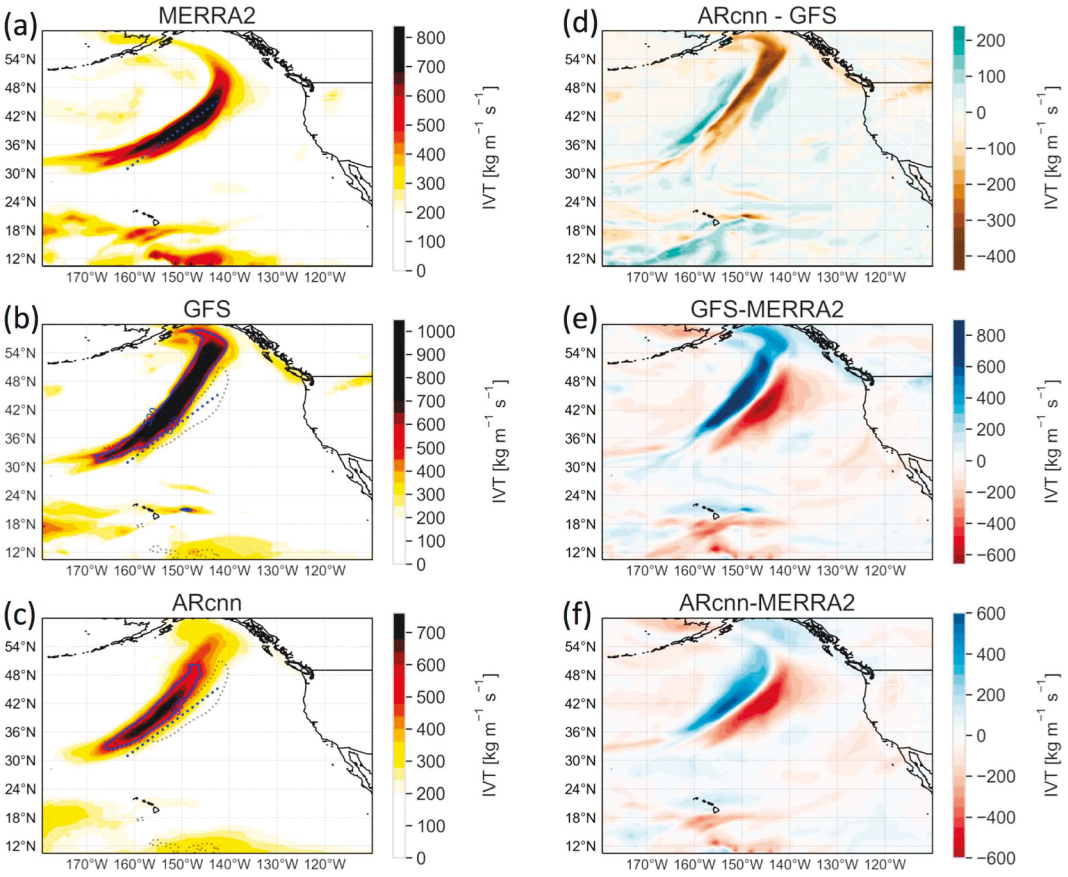

Figure 1. Forecasts and analysis valid for IVT fields on 29 November 2017. (a) MERRA-2 analysis field with the IVT = 600-kg m^{-1} s^{-1} contour (solid) and dominant storm axis (dotted) as determined by IVT > 350-kg m^{-1} s^{-1} raw image moment. (b) GFS 96-hr forecast with the MERRA-2 600 IVT contour and dominant storm axis. (c) ARcnn-IVT 96-hr forecast with the MERRA-2 600 IVT contour and dominant storm axis. (d) Difference between ARcnn-IVT and GFS. (e) Difference between GFS and MERRA-2 IVT field. (f) Difference between GFS and MERRA-2 IVT field.

2.5. ARcnn Example

ARcnn output valid for 29 November 2017 illustrates potential forecast improvements (Figure 1). The GFS forecast IVT field at 96-hr lead time (Figure 1b) is input into the 96-hr ARcnn. Once the network has been trained, a postprocessed forecast is generated within milliseconds. The IVT field passes through ARcnn (Figure S1), and a corrected field is produced (Figure 1c). The resultant field is compared against ground truth (Figure 1a). GFS overpredicts the magnitude of IVT, has a notable location error, and misses the primary orientation of the storm. After the GFS IVT field is processed with ARcnn, the network correctly reduces the magnitude of peak IVT, particularly at high latitudes near the Alaskan Coast, moving the dominant IVT signal southward and eastward (Figure 1d). Additionally, ARcnn reorients the dominant AR spatial axis to a more accurate zonal direction (Figure 1e versus 1f), leading to a more accurate forecast. Figure 1 is a representative sample of the method drawn from the top 5% of corrected 96-hr events in the testing data set (as measured by root-mean-square error (RMSE)).

3. Verification Metrics and GFS Error Patterns

Forecast error (e) is defined as the difference between the forecasted IVT field (f) and the ground truth (r) IVT field ($e = f - r$) at a given time and location. We have applied four metrics to the forecast systems: RMSE, bias (Bias), centered root-mean-square error (CRMSE), and spatial Pearson correlation (PC) coefficient. Bias and CRMSE arise from a decomposition of RMSE (Taylor, 2001). Bias represents the systematic error, defined as the mean error over the test data set (Bias = $\bar{e}$). CRMSE is the remaining random error and conditional

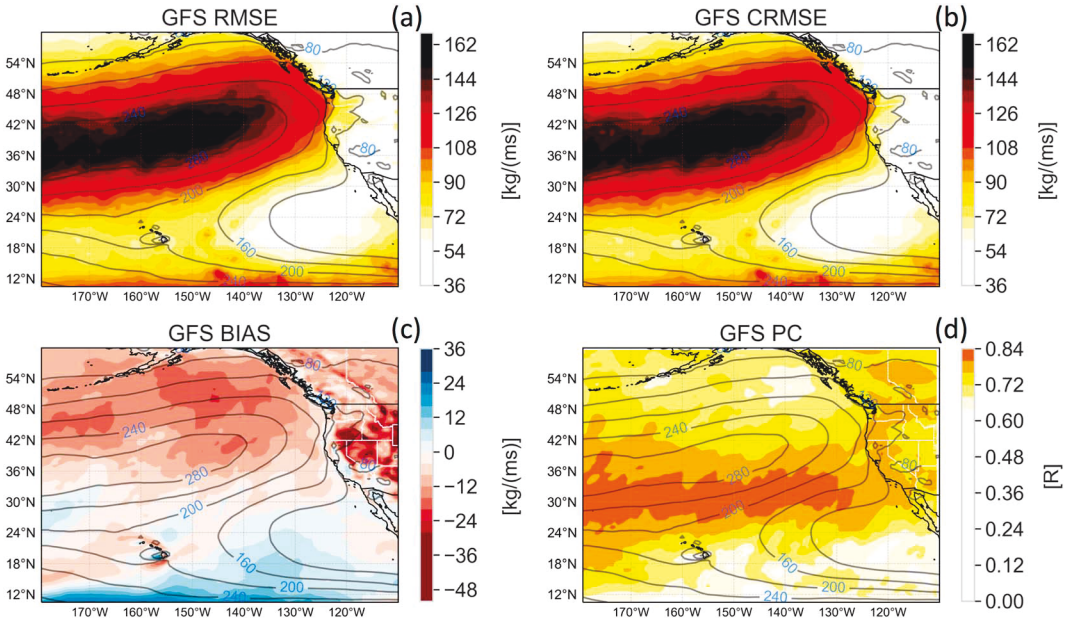

Figure 2. Spatial distribution of 96-hr forecast GFS. (a) RMSE. (b) CRMSE. (c) GFS forecast Bias. (d) Pearson correlation (in color) and (a–d) climatological AR field (in contour). Forecast dates range from October 2006 to April 2018.

biases, which contains the error not present from mean shift $\left(\mathrm{CRMSE} = \left\{\frac{1}{N}\sum_{n=1}^{N}\left[(f-\bar{f})-(r-\bar{r})\right]^2\right\}^{0.5}\right)$.

Finally, the Pearson correlation indicates the linear relationship between the forecasted and observed time series ($PC = \frac{E(f,r)}{\sigma_f \sigma_r}$).

3.1. GFS Error Patterns

The largest sources of GFS forecast error occur predominantly in the locations with high climatological IVT, indicating that AR position, magnitude, and timing constitute a large fraction of total error. Figure 2 shows the 96-hr error metrics for every GFS forecast in the data set. The RMSE field is dominated by random error and conditional bias over systematic error (as indicated by high values of CRMSE (Figure 2b), as compared with Bias (Figure 2c)). The AR corridor, defined here as the 200-kg m^{-1} s^{-1} IVT contour, for the 2006–2018 MERRA-2 climatology (Figure 2, contours), coincides with the greatest magnitude of CRMSE in the field. The model systematically underpredicts IVT magnitude at high latitudes and overpredicts IVT at low latitudes. The highest levels of PC occur on the southern flank of the AR corridor, in the climatological subtropical jet region (Figure 2d). This may be associated with the lower predictability of mesoscale frontal waves associated with ARs. Conversely, the latitudinal band of high predictability exists within the jet region and is an area of largely synoptically forced IVT processes. This latitudinal band of predictability is consistent with findings in Lavers et al. (2016).

4. Results

All statistics will be presented from a seasonal perspective derived from the testing data set (October 2017 to April 2018). The Guan and Waliser (2015) AR detection algorithm identified an AR present in 76% of the forecast periods. The AR distribution is not spatially uniform (Figure 2, contour), with a skewness toward high latitude, with landfalls predominantly in Oregon, Washington, and southern British Columbia.

ARcnn-IVT performance is evaluated at 3-hourly forecast intervals out to 12 hr, a 12-hr forecast interval out to 24 hr, and in 24-hr increments from 24-hr forecast lead until the 168-hr lead (seven days). Results for each forecast system are resampled 2,000 times for error metrics using a 30% split; the variance (color shade) of the bootstrapped sample are small compared to the mean (Figure 3). At 3-hr lead time, GFS and ARcnn-IVT outperform persistence and climatology, with the postprocessed ARcnn-IVT further improving on

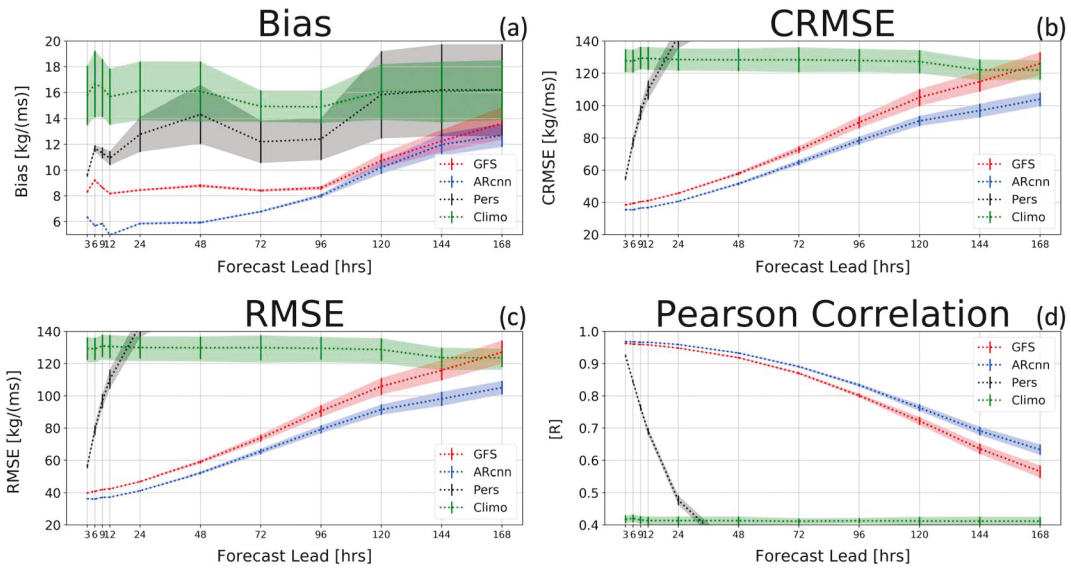

Figure 3. ROI average temporal evolution of (a) Bias, (b) CRMSE, (c) RMSE, and (d) PC of raw GFS, ARcnn, persistence (Pers), and climatology (Climo) forecasts. Resampled bootstrap variance intervals are shown for each forecast.

Bias and CRMSE over GFS. At the fifth forecast day the correction of ARcnn-IVT bias begins to deteriorate and the bias is statistically even between GFS and ARcnn-IVT after this point (Figure 3a). CRMSE (Figure 3b) continues to improve as compared to GFS for the entire testing period. Importantly, the magnitude of CRMSE dominates Bias, and therefore, the RMSE is improved (Figure 3c). At the seventh day forecast lead GFS has a larger RMSE than climatology. However, ARcnn-IVT remains the most skillful forecast (by total RMSE). The magnitude of RMSE error at the seventh day for ARcnn-IVT is equal to that of the GFS at the fifth day. This is due to the reduction of CRMSE by the postprocessing technique. Similarly, ARcnn-IVT has a higher correlation with the ground truth at every lead time, with statistically significant differences starting at hour 12. The PC of the ARcnn-IVT of the seventh day is equal to that of the sixth day for the GFS forecast.

Figure S5 shows the spatial distribution of RMSE, CRMSE, and Bias for GFS and ARcnn-IVT for the full testing data set. Figure 4 shows the spatial ARcnn-IVT metrics of performance at the 96-hr forecast lead, with cool colors indicating that ARcnn-IVT is improving the GFS forecast, conditioned on forecasted IVT values with over 250 kg m^{-1} s^{-1}, to ensure that the network is correcting for high vapor transport events. After ARcnn is applied, each pixel is assessed for RMSE, Bias, CRMSE, and PC, resampling 1,000 times utilizing 50% of the available data field in order to estimate error metrics. Importantly, RMSE (Figure 4a) at almost every grid point is decreased, indicating forecast improvement. Additionally, PC (Figure 4d) is improved at most locations with very few exceptions in the spatial domain, indicating a more skillful forecast.

The Bias field (Figure 4c) shows the least improvement, where ARcnn-IVT systematically underpredicts the magnitude of high-valued IVT. Overlaid on the figure are the $\pm$40 contours of GFS Bias. It is clear that the dominant sources of systematic error are targeted by ARcnn (as indicated by cool colors contained inside the $\pm$40 contours; Figures S5c and S5f). However, the strongest failure in the Bias field comes over the areas of coastal landfall. The field is almost uniformly improved for CRMSE (Figure 4b).

Due to a low contribution of systematic error compared to random error and conditional bias, the RMSE is still dominantly benefiting with ARcnn postprocessing. Overall, ARcnn generates an IVT field with significantly more skill than GFS. When compared to GFS, ARcnn increases correlation between ground truth and predictions at all lead times (0.5–12% increase), and the method improves RMSE at forecast leads ranging from 3 hr to seven days (9–17% reduction), equivalent to an increased forecast skill time horizon of 24- and 48-hr/day improvements, respectively. For context, NWP forecast systems, through model improvements and assimilation of additional observational data, have historically achieved an RMSE error skill improvement of approximately one day every 10 years (Magnusson & Källén, 2013).

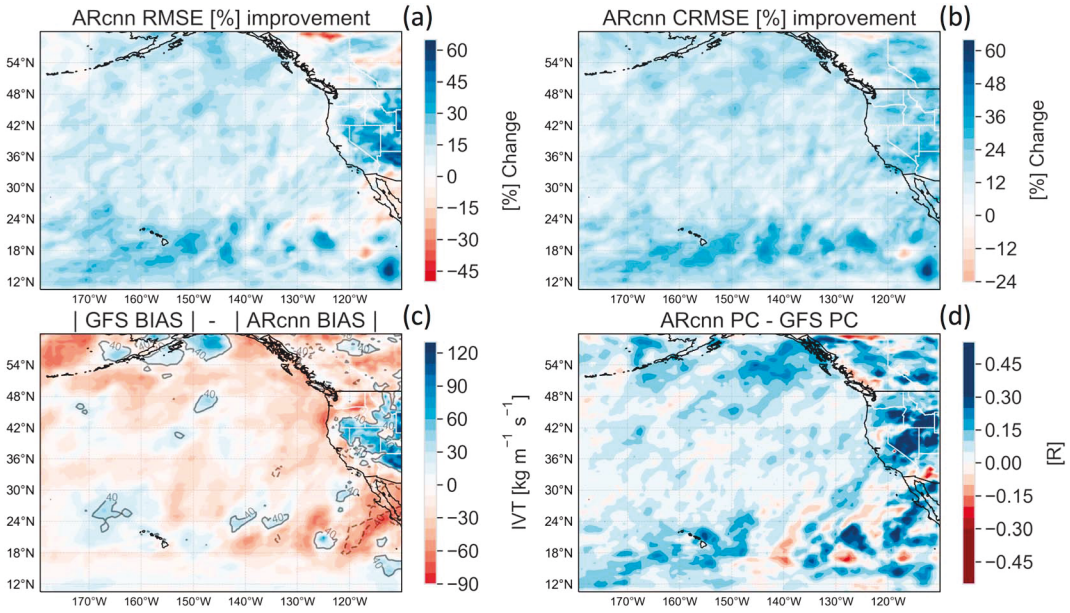

Figure 4. Spatial distribution of percent improvement of 96-hr IVT forecast after ARcnn postprocessing for (a) RMSE and (b) CRMSE. Contours indicated average IVT field. Spatial distribution of the 96-hr forecast difference between GFS minus ARcnn for (c) Bias (contours indicate GFS Bias fields of 40 units kg m^{-1} s^{-1}; dashed lines are negative) and (d) Pearson correlation. Calculated for locations when IVT forecast is over 250 kg m^{-1} s^{-1}. All dates from October 2017 to April 2018 testing data set. In all plots (a–d), cool colors imply that the CNN postprocessing is improving the forecast.

Interpretable CNNs are an active area of research in the machine learning community (e.g., Kuo et al., 2019), and ongoing research involves using CNNs to elucidate physical processes associated with forecast error. We speculate that ARcnn-IVT improvements to CRMSE involve corrections to conditional bias. Upon exhaustive inspection of individual testing forecasts, it appears that ARcnn is recognizing common IVT structures and correcting the IVT fields in similar ways given that shape. Conditional bias, that is, conditioned on storm shape and magnitude, is the most accurate terminology to describe this correction. IVT systems that appear similar to Figure 1b are similarly corrected, with a reduction in high-latitude IVT and a zonal elongation of the IVT signal. Whereas with IVT fields that are zonally stunted, ARcnn reduces the total IVT and moves the IVT signal eastward, indicating that GFS typically propagates this signal too slowly. CNNs are adept at modulating output based on input spatial field encodings. The strength of this method is the adaptive adjustment given a wide range of forecasted fields. This kind of correction results mostly in a CRMSE reduction rather than a Bias correction.

Importantly, coastal landfalling IVT 96-hr forecast RMSE is significantly improved for IVT forecasts greater than 250 kg m^{-1} s^{-1}. A detailed examination of coastal error (RMSE, CRMSE, Bias, and PC) can be found in Figures S2 and S3. The RMSE error reduction is found to be significant (90th percentile) which is important for the societal impact of landfalling ARs. Similar error reduction spatial patterns were observed for all forecast lead times (not shown). For low IVT forecasts (IVT < 250 kg m^{-1} s^{-1}; Figures S3 and S4), the improvement in forecast skill (as measured by RMSE, CRMSE, Bias, and PC) is even greater, with a significant improvement to RMSE, CRMSE, and PC, and no significant change to Bias.

5. Summary and Conclusion

This paper explored the utility of CNNs to improve IVT zero- to seven-day forecast skill. We have shown that CNNs can be used to improve forecast prediction of the GFS numerical weather prediction model for the North American West Coast and Eastern Pacific IVT 3–168-hr forecasts. This postprocessing is beneficial at every forecast lead time in reducing full-field CRMSE and improves Bias out to five forecast days, leading to a full-field RMSE improvement. ARcnn yields significantly higher PC between forecasted and ground truth values at all lead times over 12 hr. ARcnn provides a forecast that has greater skill than climatology, compared to GFS that degraded below climatological skill at seven-day lead. Ongoing work involves testing this method on an ensemble system to determine the benefit on accuracy and uncertainty quantification.

CNN postprocessing was shown here to increase IVT forecast skill. Additionally, the success of deep learning relies on the quantity of data. As forecasts are produced, CNN postprocessing techniques stand to improve as a more fully sampled distribution of AR activity is realized. CNNs continue to evolve, and model architectures are continuously under development. Opportunity exists for the weather prediction community to leverage computer vision advances. While a stand-alone machine learning weather prediction that competes with modern NWP has not been developed, combining numerical weather prediction with a data-derived CNN deep learning correction is a logical step in forecast improvement.

Acknowledgments
Data sets used to create analyses supporting this study are hosted within the University of California San Diego Library Digital Collections (https://doi.org/10.6075/J0D798R5). This study is supported by the U.S. Army Corps of Engineers (USACE)-Cooperative Ecosystem Studies Unit (CESU) as part of Forecast Informed Reservoir Operations (FIRO), grant W912HZ-15-2-0019 and the California Department of Water Resources Atmospheric River Program, grant 4600010378 TO#15 Am 22. The authors would like to thank Microsoft's AI for Earth project (https://www.microsoft.com/aiforearth) for contributing computational resources, making this work possible. Additionally, we thank the anonymous reviewers for their insightful comments that strengthened this work.

References

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends® in Machine Learning*, *2*(1), 1–127. https://doi.org/10.1561/2200000006

Bojarski, M., Yeres, P., Choromanska, A., Choromanski, K., Firner, B., Jackel, L., & Muller, U. (2017). Explaining how a deep neural network trained with end-to-end learning steers a car, 1–8. Retrieved from http://arxiv.org/abs/1704.07911

Carter, G. M., Dallavalle, J. P., & Glahn, H. R. (1989). Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Weather and Forecasting*, *4*(3), 401–412. https://doi.org/10.1175/1520-0434(1989)004<0401:SFBOTN>2.0.CO;2

Dacre, H. F., Clark, P. A., Martinez-Alvarado, O., Stringer, M. A., & Lavers, D. A. (2015). How do atmospheric rivers form? *Bulletin of the American Meteorological Society*, *96*(8), 1243–1255. https://doi.org/10.1175/BAMS-D-14-00031.1

Delle Monache, L., Nipen, T., Deng, X., Zhou, Y., & Stull, R. (2006). Ozone ensemble forecasts: 2. A Kalman filter predictor bias correction. *Journal of Geophysical Research*, *111*, D05308. https://doi.org/10.1029/2005JD006311

Delle Monache, L., Nipen, T., Liu, Y., Roux, G., & Stull, R. (2011). Kalman filter and analog schemes to postprocess numerical weather predictions. *Monthly Weather Review*, *139*(11), 3554–3570. https://doi.org/10.1175/2011mwr3653.1

Dettinger, M. D., Ralph, F. M., Das, T., Neiman, P. J., & Cayan, D. R. (2011). Atmospheric rivers, floods and the water resources of California. *Water*, *3*(2), 445–478. https://doi.org/10.3390/w3020445

Dong, C., Loy, C. C., He, K., & Tang, X. (2014). *Learning a deep convolutional network for image super-resolution*, *Lecture Notes in Computer Science*, (pp. 184–199). https://doi.org/10.1007/978-3-319-10593-2_13

Dueben, P. D., & Bauer, P. (2018). Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, *11*(10), 3999–4009. https://doi.org/10.5194/gmd-11-3999-2018

Gagne, D. J., McGovern, A., Haupt, S. E., Sobash, R. A., Williams, J. K., & Xue, M. (2017). Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Weather and Forecasting*, *32*(5), 1819–1840. https://doi.org/10.1175/waf-d-17-0010.1

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., et al. (2017). The Modern-Era Retrospective Analysis for Research and Applications, version 2 (MERRA-2). *Journal of Climate*, *30*(14), 5419–5454. https://doi.org/10.1175/JCLI-D-16-0758.1

Gershunov, A., Shulgina, T., Ralph, F. M., Lavers, D. A., & Rutz, J. J. (2017). Assessing the climate-scale variability of atmospheric rivers affecting western North America. *Geophysical Research Letters*, *44*, 7900–7908. https://doi.org/10.1002/2017GL074175

Glahn, H. R., & Lowry, D. A. (1972). The use of Model Output Statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology*, *11*(8), 1203–1211. https://doi.org/10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2

Guan, B., & Waliser, D. E. (2015). Detection of atmospheric rivers: Evaluation and application of an algorithm for global studies. *Journal of Geophysical Research: Atmospheres*, *120*, 12,514–12,535. https://doi.org/10.1002/2015JD024257

Hacker, J. P., & Rife, D. L. (2008). A practical approach to sequential estimation of systematic error on near-surface mesoscale grids. *Weather and Forecasting*, *22*(6), 1257–1273. https://doi.org/10.1175/2007waf2006102.1

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. Proceedings of CVPR, 19, 770–778. https://doi.org/10.1016/0141-0229(95)00188-3

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527

Homleid, M. (1995). Diurnal corrections of short-term surface temperature forecasts using the Kalman filter. *Weather and Forecasting*, *10*(4), 689–707. https://doi.org/10.1175/1520-0434(1995)010<0689:DCOSTS>2.0.CO;2

Kingma, P., Diederik, B., & Lei, J. (2014). Adam: A method for stochastic optimization. ArXiv.

Kuo, C. C. J., Zhang, M., Li, S., Duan, J., & Chen, Y. (2019). Interpretable convolutional neural networks via feedforward design. *Journal of Visual Communication and Image Representation*, *60*, 346–359. https://doi.org/10.1016/j.jvcir.2019.03.010

Kurth, T., Treichler, S., Romero, J., Mudigonda, M., Luehr, N., Phillips, E., et al. (2018). Exascale deep learning for climate analytics. Retrieved from http://arxiv.org/abs/1810.01993

Lamjiri, M. A., Dettinger, M. D., Ralph, F. M., & Guan, B. (2017). Hourly storm characteristics along the U.S. West Coast: Role of atmospheric rivers in extreme precipitation. *Geophysical Research Letters*, *44*, 7020–7028. https://doi.org/10.1002/2017GL074193

Lavers, D. A., Waliser, D. E., Ralph, F. M., & Dettinger, M. D. (2016). Predictability of horizontal water vapor transport relative to precipitation: Enhancing situational awareness for forecasting western U.S. extreme precipitation and flooding. *Geophysical Research Letters*, *43*, 2275–2282. https://doi.org/10.1002/2016GL067765

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2015.7298965

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, *20*(2), 130–141. https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2

Magnusson, L., & Källén, E. (2013). Factors influencing skill improvements in the ECMWF forecasting system. *Monthly Weather Review*, *141*(9), 3142–3153. https://doi.org/10.1175/mwr-d-12-00318.1

McCarty, W., Coy, L., Gelaro, R., Huang, A., Merkova, D., Smith, E. B., et al. (2016). MERRA-2 input observations: Summary and assessment. *Technical Report Series on Global Modeling and Data Assimilation*, *46*(October). Retrieved from). https://gmao.gsfc.nasa.gov/pubs/docs/McCarty885.pdf

McCollor, D., & Stull, R. (2008). Hydrometeorological accuracy enhancement via postprocessing of numerical weather forecasts in complex terrain. *Weather and Forecasting*, *23*(1), 131–144. https://doi.org/10.1175/2007WAF2006107.1

McGovern, A., Elmore, K. L., Gagne, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., et al. (2017). Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bulletin of the American Meteorological Society*, *98*(10), 2073–2090. https://doi.org/10.1175/bams-d-16-0123.1

Moorthi, S., Pan, H.-L., & Caplan, P. (2001). Changes to the 2001 NCEP Operational MRF/AVN Global Analysis/Forecast System. *NWS Technical Procedures Bulletin*, *484*, 1–14.

Nielsen, M. A. (2015). *Neural networks and deep learning*. Determination Press.

Ralph, F. M., Neiman, P. J., & Wick, G. A. (2004). Satellite and CALJET aircraft observations of atmospheric rivers over the eastern North Pacific Ocean during the winter of 1997/98. *Monthly Weather Review*, *132*(7), 1721–1745. https://doi.org/10.1175/1520-0493(2004)132<1721:SACAOO>2.0.CO;2

Ralph, F. M., Rutz, J. J., Cordeira, J. M., Dettinger, M., Anderson, M., Reynolds, D., et al. (2018). A scale to characterize the strength and impacts of atmospheric rivers. *Bulletin of the American Meteorological Society*, *100*(2), 269–289. https://doi.org/10.1175/bams-d-18-0023.1

Rasp, S., & Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, *146*(11), 3885–3900. https://doi.org/10.1175/MWR-D-18-0187.1

Richardson, L. F. (1922). *Weather prediction by numerical process*. Cambrige University Press.

Roeger, C., Stull, R., McClung, D., Hacker, J., Deng, X., & Modzelewski, H. (2003). Verification of mesoscale numerical weather forecasts in mountainous terrain for application to avalanche prediction. *Weather and Forecasting*, *18*(6), 1140–1160. https://doi.org/10.1175/1520-0434(2003)018<1140:vomnwf>2.0.co;2

Scher, S. (2018). Toward data-driven weather and climate forecasting: Approximating a simple general circulation model with deep learning. *Geophysical Research Letters*, *45*, 12,616–12,622. https://doi.org/10.1029/2018GL080704

Scher, S, & Messori, G. (2019). Weather and climate forecasting with neural networks: Using GCMs with different complexity as study-ground. *Geoscientific Model Development Discussions*, (March), 1–15. https://doi.org/10.5194/gmd-2019-53

Scher, S., & Messori, G. (2018). Predicting weather forecast uncertainty with machine learning. *Quarterly Journal of the Royal Meteorological Society*, *144*(717), 2830–2841. https://doi.org/10.1002/qj.3410

Schulzweida, U. (2019). CDO user guide (version 1.9.6). https://doi.org/10.5281/zenodo.2558193

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W., & Woo, W. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting, 1–12. Retrieved from http://arxiv.org/abs/1506.04214

Sodemann, H., & Stohl, A. (2013). Moisture origin and meridional transport in atmospheric rivers and their association with multiple cyclones*. *Monthly Weather Review*, *141*(8), 2850–2868. https://doi.org/10.1175/mwr-d-12-00256.1

Stensrud, D. J., & Skindlov, J. A. (2002). Gridpoint predictions of high temperature from a mesoscale model. *Weather and Forecasting*, *11*(1), 103–110. https://doi.org/10.1175/1520-0434(1996)011<0103:gpohtf>2.0.co;2

Stensrud, D. J., & Yussouf, N. (2003). Short-range ensemble predictions of 2-m temperature and dewpoint temperature over New England. *Monthly Weather Review*, *131*(10), 2510–2524. https://doi.org/10.1175/1520-0493(2003)131<2510:sepomt>2.0.co;2

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. https://doi.org/10.1175/JHM-D-15-0075.1

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*, *106*(D7), 7183–7192. https://doi.org/10.1029/2000JD900719

Toms, B. A., Kashinath, K. P., & Yang, D. (2019). Deep learning for scientific inference from geophysical data: The Madden-Julian Oscillation as a test case. Retrieved from http://arxiv.org/abs/1902.04621

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th international conference on Machine learning – ICML'08 (pp. 1096–1103). New York, New York, USA: ACM Press. https://doi.org/10.1145/1390156.1390294

Warner, M. D., Mass, C. F., & Salathé, E. P. (2012). Wintertime extreme precipitation events along the Pacific Northwest Coast: Climatology and synoptic evolution. *Monthly Weather Review*, *140*(7), 2021–2043. https://doi.org/10.1175/mwr-d-11-00197.1

Wick, G. A., Neiman, P. J., Ralph, F. M., & Hamill, T. M. (2013). Evaluation of forecasts of the water vapor signature of atmospheric rivers in operational numerical weather prediction models. *Weather and Forecasting*, *28*(6), 1337–1352. https://doi.org/10.1175/WAF-D-13-00025.1

Wilks, D. S., & Hamill, T. M. (2007). Comparison of ensemble-MOS methods using GFS reforecasts. *Monthly Weather Review*, *135*(6), 2379–2390. https://doi.org/10.1175/mwr3402.1

Zhang, K., Zuo, W., Chen, Y., Meng, D., & Zhang, L. (2017). Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, *26*(7), 3142–3155. https://doi.org/10.1109/TIP.2017.2662206

Zhu, Y., & Newell, R. E. (1998). A proposed algorithm for moisture fluxes from atmospheric rivers. *Monthly Weather Review*, *126*(3), 725–735. https://doi.org/10.1175/1520-0493(1998)126<0725:APAFMF>2.0.CO;2

References From the Supporting Information

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2016). TensorFlow: Large-scale machine learning on heterogeneous distributed systems. Retrieved from http://arxiv.org/abs/1603.04467

Chollet, F. (2015). Keras. GitHub repository. Retrieved from https://github.com/fchollet/keras

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. Retrieved from http://arxiv.org/abs/1502.03167

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. Retrieved from http://arxiv.org/abs/1511.07122