Bulletin of Atmospheric Science and Technology (2020) 1:373–385
https://doi.org/10.1007/s42865-020-00028-z

RESEARCH ARTICLE

# Post-processing rainfall in a high-resolution simulation of the 1994 Piedmont flood

Scott Meech[1] · Stefano Alessandrini[1] · William Chapman[2] · Luca Delle Monache[2]

Received: 7 April 2020 / Accepted: 2 December 2020/ Published online: 3 January 2021

**Abstract**
In November 1994, a catastrophic flooding event occurred in the Piedmont region in Northwestern Italy over a period of about 3 days. The large time and spatial scales associated with this event prompted a number of reanalysis studies to assess the forecast skill of the models. This paper investigates another forecasting technique using the Weather Research and Forecasting (WRF) model coupled with post-processing techniques: the analog ensemble (AnEn) and the convolutional neural network (CNN). The complex topography found in this region presents a challenge for numerical weather prediction (NWP) models especially for events such as these, where the orography is crucial in determining the distribution and amount of precipitation. By applying these post-processing techniques to WRF model output, significant improvements were observed in the accumulated precipitation fields during the flooding event in both techniques, although improvements using the CNN were at the expense of underestimating the highest precipitation.

**Keywords** Analog Ensemble · Convolutional neural network · Piedmont flood

## 1 Introduction

The 1994 Piedmont flood was characterized by 72-h (4–6 November 1994) continuous rainfall over the Italian region with recorded amounts of about 600 mm of accumulated precipitation in some locations. Several rivers and creeks in the Po valley devastated cities like Alessandria with significant damage to the infrastructure and a total of 70 casualties in the region. Several numerical studies (Ferretti et al. 2000; Rotunno 2001; Cassardo et al. 2002; Buzzi et al. 1998) showed that the flow interaction with the orography determined the extraordinary rainfall

✉ Scott Meech
smeech@ucar.edu

[1] National Center for Atmospheric Research (NCAR), 3450 Mitchell Ln, Boulder, CO 80301, USA

[2] Scripps Institute of Oceanography, University of California, La Jolla, CA, USA

amount and its spatial pattern. For a thorough meteorological description of the event, we direct the reader to Binder (1996), Doswell et al. (1998), and Romero et al. (1998).

In this work, we want to evaluate the performances of high-resolution numerical prediction models (NWP) such as the Weather Research and Forecasting model (WRF) (Skamarock et al. 2008) coupled with post-processing techniques based on the analog ensemble (AnEn) (Delle Monache et al. 2013) and convolutional neural networks (CNN) (Nielsen 2015) in reconstructing the precipitation amounts during the 1994 flood event. It is worth noting that all these models (WRF, AnEn, and CNN) were not available in 1994. In fact, they have been developed by the scientific community in the following years and, also, the computational power available in 1994 would not have allowed the use of NWP at a high-resolution (~3 km) for operational forecasting in most of the European national weather services. The first novel aspect of this study is to highlight the level of performances that could be achieved if these models were adopted back in 1994 and, similarly, if such an event should repeat in the future.

The WRF model (Skamarock et al. 2008) is a popular atmospheric model supported and developed by the National Center for Atmospheric Research (NCAR) and community contributions. Since its release in 2000, WRF has become one of the most used atmospheric models in terms of registered users and publications. The model provides a range of Earth system prediction applications, such as air chemistry, hydrology, wildland fires, hurricanes, and regional climate (Powers et al. 2017). WRF is a non-hydrostatic model that fully conserves mass and includes several options of dynamic cores, physical parameterizations, and nesting designed for the 1–10-km grid scale for research applications and operational forecasting. The last version (4.1.5) was released on March 10, 2020. The model version used in this study is version 3.5.1 (released September 23, 2013) which is the model version supported by the Global Climatology Analysis Tool (GCAT) (NCAR 2020; Alessandrini et al. 2017) used to run the climatology simulations. In this work, we have used WRF to generate 30-year 24-h accumulated precipitation fields at 3-km resolution for the months of October and November over the Italian Piedmont region.

The AnEn approach is an analog-based technique that has been used as a post-processing scheme and therefore can be considered a model output statistic approach (MOS) for NWP systems (e.g., Hamill and Whitaker 2006; Delle Monache et al. 2013; Sperati et al. 2017). As an additional feature when compared with most of the MOS described in the literature, the AnEn can provide an ensemble or probabilistic forecast of any variable for which an archive of past observations or analysis values is available. Being that the AnEn in its standard implementation is based only on deterministic forecasts, it can generate an ensemble with lower computational effort than what is required for running an NWP dynamical ensemble. Also, the analog ensemble searches for analog forecasts in an archive dataset and uses the correspondent verifying observations to build the ensemble forecast. So, the sampling of (analog) ensemble members is made from the observed distribution, and the corresponding "true" probability density function is estimated. The AnEn has been used in a wide range of applications ranging from meteorological predictions (Delle Monache et al. 2013; Sperati et al. 2017), tropical cyclone intensity forecasts (Alessandrini et al. 2018), air quality predictions (Delle Monache et al. 2020), and renewable energy forecasting (Alessandrini et al. 2015). The works that are more strictly related to the current one are Hamill and Whitaker (2006) and Keller et al. (2017). In the former, the technique is used for precipitation reforecasts generated with a coarse T62 (spectral resolution) version of NCEP's Global Forecast System. The AnEn showed some potential to improve precipitation forecasts even though the level of performance was strongly dependent on the calibration with respect to the set of predictors and neighborhood size. In the latter, the AnEn is used to downscale precipitation estimates from a regional reanalysis of 6.2-km for Europe which was providing the predictors. Some parameters of AnEn, such as the choice of predictors or the ensemble size, were tuned to optimize the performances. In this work, we have used the AnEn to improve the precipitation estimate from WRF over the Piedmont region during the 1994 flood event.

Computational hardware advances, (i.e., graphical processing units), and newly developed flexible software packages (i.e., Chollet 2015) have inspired recent growth in the field of deep neural networks (DNNs). DNNs are machine-learning algorithms that consist of several layers of interconnected nodal points that are activated by nonlinear functions between each layer (Nielsen 2015). It is the ability of the DNNs to map nonlinear relationships that make them unique among MOS systems. They have begun to be used as NWP post-processing methods and have been shown to outperform traditional MOS and ensemble MOS methods (i.e., Lauret et al. 2014; Rasp and Lerch 2018). CNNs, DNNs which process matrix data rather than nodal data, expand the spatial scope of the input forecast field by examining non-local spatial information within the network. CNNs have been making large advances in the atmospheric science community and have been used to construct complete (although simplified) weather prediction models (i.e., Dueben and Bauer 2018; Scher 2018; Weyn et al. 2019), make atmospheric physics discovery (i.e., Toms et al. 2019; Gagne Ii et al. 2019), and post-process NWP weather forecasts (i.e., Chapman et al. 2019). Additionally, CNNs have been used to improve precipitation estimation through spatial downscaling (Hsu et al. 2019), and precipitation nowcasting (Kim et al. 2017). The addition of spatial encoding in CNN MOS systems adds spatial information not accessed by traditional MOS methods and can therefore encode non-local weather patterns which aids with the improvement of conditional biases (Chapman et al. 2019). Like for the AnEn, CNN is also used in this work to improve the precipitation estimate from WRF over the Piedmont region during the 1994 flood event.

As a consequence, the second novel aspect of this study is to compare a point-based MOS (AnEn) with a spatial-based MOS (CNN). In fact, the AnEn as implemented in this work is independently trained only from past forecasts and observations at each grid point. The CNN, as already mentioned, is trained with past forecasts and observations from the entire grid. Hence, the correction over a single-grid point can benefit from additional information from the spatial patterns of the biases rather than being based on the local biases only.

This paper is organized as follows: in Section 2, we describe the observation dataset, and in Section 3, the meteorological simulations used to build the training dataset. In Section 4, the two post-processing techniques (AnEn and CNN) are described. In Section 5, the performances of WRF, AnEn, and CNN are compared with the conclusions drawn in Section 6.

## 2 Observation dataset

Observation data were derived from two measurement networks located in Northwestern Italy: the current telemeasure network of the regional meteorological service ARPA Piemonte (1988—present) and the Italian legacy system “Servizio Idrografico e Mareografico Nazionale” mechanical and manual station network (SIMN) from 1913 to 2002. These combined measurement networks are maintained by the Forecasting Systems Department of ARPA Piemonte and referred to as the NWIOI dataset (Ronchi et al. 2008). Both of the sensor networks were subject to quality control measures. Logical checks monitored the consistency of the data and excluded physically impossible values; a station’s series of measurements were compared with surrounding stations, as well as checks for anomalous persistence of a single value. The number of stations increased dramatically from 119 gauges in the 1950s to 386 in 2009. During the 1994 flooding event, there were about 200 rain gauge stations cumulatively collecting data between these two networks.

These data sets were interpolated onto a regular grid using an optimal interpolation technique (Arpa 2010) due to the increasing number and proximity of the stations over the total collection period from 1957 to 2009. A Voronoi Tessellation was carried out yearly on active stations to estimate the average two-dimensional radius based on the Tessellation cell and an approximate circular surface of the cell, which is indicative of the average distance between the stations throughout the analysis period. This technique makes it possible to produce a final estimate that is not influenced by extreme values as the number of stations increases in the analysis. The ultimate result was a gridded data set of temperature and rainfall values in an optimal 15-km (0.125 °) resolution which best represented the full temporal period of collections. The domain extends from 44.0 to 46.5 N and 6.5–9.5 E with 297 grid points residing within the Piedmont administrative territory.

## 3 Meteorological simulations

The WRF model was run on GCAT, whose development has been supported by the US Army (Alessandrini et al. 2017), from 1984 to 2013 for a period of 2 months (October 1 00Z–November 30 00Z) each year using initialization and lateral boundary conditions from the NOAA (National Oceanic and Atmospheric Administration) Climate Forecast System Reanalysis (CFSR) data set and nudged towards the Global Data Assimilation System (GDAS) observations. WRF is run with 3 domains centered at 44.909 N, 8.610 E, downscaling the CFSR 0.5° data set from 30-, 10-, to 3.3-km resolution on the innermost domain and 37 layers in the vertical up to 50 mb. The model is restarted every 5 days at 00Z and leverages 6-hourly updated boundary conditions. The WRF ARW (Advanced Research WRF) core used WSM (WRF single moment) 6-class graupel scheme microphysics, RRTM (rapid radiative transfer model) and Dudhia long- and short-wave radiation schemes, YSU (Yonsei University) planetary boundary layer (PBL) scheme, and the Noah land-surface model.

Precipitation data is accumulated from each hour of WRF model output from 12 to 12 UTC in order to match the accumulated daily precipitation fields from the observations temporally. The summed daily WRF precipitation totals are then extracted from the nearest WRF grid point for each gridded observation data point and stored. The 24-h mean of direct WRF variables U10, V10, surface pressure, and 2-m-temperature are extracted along with relative humidity, geopotential height, U, V, and temperature interpolated using WRF-Python (Ladwig 2017) to pressure levels 850, 700, and 500 mb and used as input to the post-processing algorithms, AnEn and CNN.

Given that WRF is driven by reanalysis fields as initial and boundary conditions, the simulations cannot be considered as forecasts but rather downscaled reanalysis. In fact, in a forecast simulation, the initial conditions are usually provided by an analysis model, and the boundary conditions are forecasts from a global forecast model like the NOAA global forecasting system (GFS) for instance. However, given the accuracy achieved in the recent years by global model forecasts in the 0–24-h range, the performance of our WRF simulations can be considered very similar to those achievable by running a 0–24-h forecast. Also, our approach guarantees to build a 30-year training dataset with a fixed model configuration and a stable model error probability density function (PDF). Having a fixed NWP configuration over a long period is of great advantage when applying any post-processing technique to correct the NWP output. In terms of the comparison between CNN and AnEn, we believe that the outcomes of this study are very similar to those we would obtain if an archive of 0–24-h forecast was used.

## 4 Post-processing techniques

The 24-h accumulated precipitation (AP) values computed by WRF and AP observations cover a 30-year period (1984–2013) for the months of October and November and are used to build the dataset used for the post-processing techniques described hereafter. As a training dataset, we have used all the years of the 30-year period except for 1994 that is used as a verification dataset. Even though the AP post-processed estimates are generated for all of the 61 days of October–November 1994, in the following sections, we focus over the period from 12 UTC of November 3 to 12 UTC of November 7 during which the highest precipitation amounts were recorded in Piedmont.

### 4.1 The analog ensemble

In this section, we briefly describe the AnEn algorithm introduced by Delle Monache et al. (2013) for 10-m wind speed and 2-m temperature ensemble predictions and already applied by Keller et al. (2017) to downscale precipitation reanalysis datasets. The AnEn exploits a dataset of point forecasts generated by an NWP model and a record of observations at the same point. In this work, we focus on AP which is the AnEn predict and (the variable to be predicted). The dataset of forecasts from WRF, besides AP, provides a set of meteorological variables used as additional predictors for the analog search. In the current work, we have used AP, the 24-h mean of 2-m temperature (T-2 m), 10-m U and V wind components (U-10, V-10), 700 mb U and V wind components (U700, V700), the product of the vertical wind velocity and relative humidity at 700 mb (W700 × RH700) and temperature at 700 mb (T700). We selected these variables because they were considered to have some relationship with the precipitation computed by WRF. Except for precipitation itself, which of course is the main predictor, all the wind components (U-10 and V-10, U700 and V700) are important to determine the orographic lifting. W700 × RH700 is also important because it can highlight orographic lifting occurring concurrently with vapor condensation. T700 and T-2 m are helpful to determine the type of precipitation (snow or rain). We did not use an objective method for the predictor's selection. On the other hand, as explained hereafter, we performed a "brute force" optimization to objectively assess the predictors' importance (among those selected). Indeed, we understand that we might have missed some variables which could also have led to some further improvement in the precipitation post-processing.

In the AnEn construction, the predictors are used to detect a given number of forecasts in the training dataset similar to the target forecast (the forecast we aim at improving) in the testing dataset. The corresponding past concurrent verifying AP observations form an ensemble forecast. In this work, we focus on improving the deterministic AP predictions. Hence, we have used the ensemble mean to generate a single value from the ensemble. The AnEn aims at detecting cases in the training dataset when the model error PDF (the distribution of the differences between predicted and observed AP) is similar to the PDF of the target forecast in the verification dataset. If such cases are found, then the verifying observations in the verification dataset and target forecast are sampled from the same PDF.

The degree of similarity between an AP forecast at a given date *t* and location to the forecasts in the training dataset at the same location is assessed by computing the distance (*Dt*), which is:

$$D_{P,t,ta} = \sum_{P} w_P D_{t,ta}, \tag{1}$$

where

$$D_{t,ta} = \sqrt{\left(P_t - P_{ta}\right)^2 + \left(P_{t-24h} - P_{ta-24h}\right)^2 + \left(P_{t+24h} - P_{ta+24h}\right)^2}, \tag{2}$$

and subscripts *t* and *ta* represent, respectively, the date of the target forecast and of a potential analog forecast in the training dataset. In Eqs. (1) and (2), *P* is the value of the predictor normalized by its standard deviation computed over the training dataset, and $w_P$ is the weight assigned to each predictor. This distance is computed over a time interval of 3 consecutive days (+ or − 24 h) to also consider similar daily trends in the predictions. As in (Alessandrini et al. 2019), a bias correction for rare events is applied when the target AP forecast exceeds 20 mm. The 20-mm threshold has been set after some tuning was carried out over the training dataset (not shown here). Also, in (Delle Monache et al. 2013), the weights $w_P$ were each specified as equal to 1, thus assigning the same importance to each predictor. Several subsequent works by Alessandrini et al. (2015) and Junk et al. (2015) have demonstrated that a brute-force weights' optimization (which is computationally feasible with a limited number of predictors, as in the current study) can improve the AnEn performance. Therefore, we have performed a weight optimization independently at each grid point by choosing the combination that minimizes the root mean squared error (RMSE) over the training dataset. Since only eight predictors are used, eight corresponding weights can be set. All the possible combinations defined with the constraint $\sum_{P=1}^{8} w_P = 1$, where $w_P \in$ [0,0.1,0.2, …,1], are tested for the AnEn prediction over the training dataset using a leave-one-out approach over the training period. Specifically, for each day, the AnEn predictions are issued for all possible combinations of weights using all the remaining days in the training for the analog search. Therefore, for any grid point of observation dataset, each weight combination can be evaluated in terms of RMSE, selecting the best one (lowest RMSE) as previously mentioned.

In Fig. 1, the mean weight for each predictor computed across all the grid points is depicted. As expected, AP gets the highest weight among the 8 predictors tested that all end up receiving, on average, positive values. In general, it means that they are all used to select analogs in the training dataset. The second predictor by importance is V700 receiving an average weight of about 0.2. This weight is consistent with the meridional wind component determining the extent of the orographic lifting of the wet air masses coming from the Mediterranean Sea from the South of the Piedmont region. Especially in the case of meridional winds, the interaction of these air masses with Ligurian Alps (South of Turin and Alessandria), and with the Alps (North of Turin and Alessandria), is enhanced. The amount of precipitation and its spatial distribution over the Piedmont region is strongly dependent on V700 as also highlighted by Rotunno (2001).

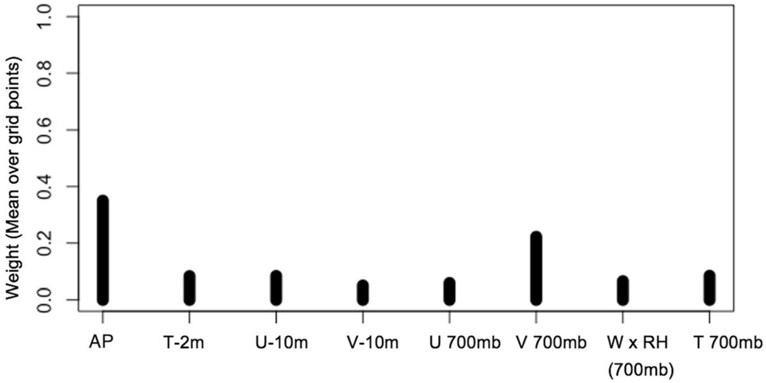

**Fig. 1** Mean over all the grid points of the weights received by each predictor after the brute force optimizations

### 4.2 Convolutional neural networks

Artificial neural networks (ANN) are capable of approximating nonlinear functions and processes (Nielsen 2015). This approximation is accomplished through a series of feed-forward matrix operations, which pass input variables through a series of "hidden" model layers, to a specified output layer. Each layer is described by the number of nodal points in that layer, with the initial layer being the number of input variables. Nodes from adjacent model layers are connected via model weights. The hidden nodal point values are determined by the sum of the product of associated model weights and the input values from the previous layer. Each nodal point is then "activated" by a nonlinear function (e.g., rectified linear unit ($ReLU$) $max(0, x)$), before passing the variables to the following layer.

Convolutional neural networks (CNN) are a subcategory of ANNs which operate on image data by training small matrix "kernel weights ($w$)" which pass over the images and distinguish relevant image features rather than nodal points; each layer operation thus results in a set of salient feature maps which are fed to the following layer of the CNN. The task of training a CNN (ANN) is to discover the optimal convolutional kernel (nodal) weights. The weights are learned iteratively through backward optimization and gradient descent, in which each iteration seeks to minimize the cost of a specified loss function ($L$) (e.g., mean squared error between the CNN output and the target post-processed variable), by determining the gradient field of the weights ($\frac{dw}{dL}$) and taking a small step in the direction opposite this gradient. For further exploration of CNNs, the reader is directed to Nielsen (2015).

Chapman et al. (2019) showed that CNNs are effective post-processing tools to develop spatial error relationships and correct bias and conditional biases within a forecast of the integrated vapor transport field. Here, we adopt a similar method but include multivariate temporal data in the convolution, and we target precipitation. We input the forecasted variables into the CNN at 3 consecutive days, in 169 WRF locations, utilizing 9 input forecast variables (AP, U-10, V-10, and T-2 m, relative humidity, U, V, W, and temperature at 700 mb) as our input, thus forming a 3 × 169 × 9 variable matrix, which forces a spatio-temporal relationship to be learned by the convolutional kernels. After two convolutional layers, the salient feature maps are then flattened and fed into a feed-forward ANN which outputs the post-processed precipitation forecasts at 169 locations for each of the 3 days (507 output nodes). Observations are used to determine the loss, as measured by the mean squared error, between the network output and the observed precipitation. The data description and train/validate/test split can be seen in Tables 1 and 2, the data category split is separated by year to allow for temporal decorrelation between training and testing samples. This data split ensures that model validation and tuning are done on representative data (validate), and the final model is tested on independent data (testing) to combat false skills associated with model overfitting.

Table 2 shows the architecture and parameters of the CNN utilized in this study. This architecture was selected after an extensive hyperparameter search, and the minimum error, as calculated on the validation data set (Table 2), determined the final network parameters. The CNN utilizes an Adam optimizer (Kingma and Ba 2015) with a learning rate that decreased from 0.001 to $5e^{-6}$ upon validation plateaus of 5 or more epochs and a batch size of 30. The network is trained until validation loss plateaus or increases for 10 or more epochs, with the final configuration training for 30 epochs and the final model weights determined by the lowest value of validation error. We note that, during training, the validation error showed no sign of overfitting. Final training, testing, and validation datasets show similar magnitudes of error. The CNN was designed and trained using the Keras (Chollet 2015) with the TensorFlow Backend (Abadi et al. 2016) python libraries.

## 5 Verification

Accumulated daily precipitation totals between the WRF model and the gridded observations averaged over a period of 30 years display a topographical bias in the Piedmont region. Complex terrain along the Italian Alps on the Northern and Western border and the Ligurian Apennines to the South typically receive an increased amount of precipitation than the lower-lying interior territory. The WRF model correlates well with the gridded observations exhibiting this same pattern (Fig. 2).

Over a period of 30 years from October 1 12Z to November 29 12Z, mean daily precipitation differences between the WRF model and the gridded observations display a slight overprediction within the low-lying interior region while areas of more complex topography show discrepancies with sharper contrast (Fig. 3). The largest departures from the observed amount can be seen in the RMSE of the Northern portion of the Piedmont region and the Southern foothills roughly spanning from Cuneo to East of Genoa.

In Fig. 4, we present the WRF, AnEn, and CNN in terms of bias during the flooding event between November 3 12Z and 7 12Z, 1994. The WRF model displayed a similar pattern of overpredicting the precipitation in the Northern and Southern portions of the Piedmont region and conversely underpredicting in the Southeast (Fig. 4). Within the interior, North–South streaks of under and over predictions occur which may be due to the timing and position of the meridional flow. The largest over predictions occurred along the extreme Northern portion of the domain in the Alps and along the South and Southwestern domain border. The largest underpredictions occurred within the aforementioned North-South streaks leading into the Alps and in the Southeast just North of the foothills around Alessandria.

**Table 1** Dataset used for CNN

| | Train | Validate | Test |
|---|---|---|---|
| Years (Oct 1–Nov 31) | 1984–2012 (excluding 1994) | 2013 | 1994 |
| Variables | Precipitation, T-2 m, U10, V10, RH700, U700, V700, W700, T700 | | |

**Table 2** Architecture of the convolutional neural network used for post-processing precipitation. Input is 3 forecast time steps, in 169 WRF locations, utilizing 9 input forecast variables (precipitation, U10, V10, and 2-m-temperature, relative humidity, U, V, W, and temperature at 700 mb). Leaky ReLU activation is utilized in every layer but the final, which is activated via a linear function for the final precipitation metric. The loss is determined by the mean squared error of the output node and the observed precipitation

| Layer number | Type | Input size | Output size | Parameters |
|---|---|---|---|---|
| 1 | Input | 3 × 169 × 9 | 3 × 169 × 9 | – |
| 2 | Conv2D | 3 × 169 × 9 | 3 × 169 × 6 | N.filt (6); kernel (5, 5); activation: leaky ReLU (alpha = 0.2); batch norm.; dropout (0.2) |
| 3 | Max pooli-ng | 3 × 169 × 6 | 1 × 84 × 6 | Pool size: (2, 2) |
| 4 | Conv2D | 1 × 84 × 6 | 1 × 84 × 12 | N.filt (6); kernel (5, 5); activation: leaky ReLU (alpha = 0.2); batch norm.; dropout (0.2) |
| 5 | Flatten | 1 × 84 × 12 | 1 × 512 | - |
| 6 | Dense | 1 × 512 | 1 × 100 | Nodes (100); activation: leaky ReLU (alpa = 0.2); batch norm |
| 7 | Dense | 1 × 100 | 1 × 507 | Nodes (507): activation: linear |
| 8 | Output | 1 × 507 | 1 × 507 | Loss (mean squared error); Adam optimizer (0.001) |

The AnEn reduced overpredictions from WRF in the extreme Northern portion of the domain in the higher elevations of the Italian Alps while exacerbating underpredictions just South of this in this area leading into the southern foothills of the Alps in the Northwestern portion of the Piedmont Valley. WRF overpredictions in the Southern Piedmont region and underpredictions to around Alessandria were reduced to a lesser extent. The interior North-South streak patterns remain largely unchanged.

The convolutional neural network simulation removed the most significant overpredictions and replaced them with a large swath of underpredictions in the Alps and meridionally through the interior region. The largest underpredictions remain in the foothills leading into the Alps to the North while the underpredictions in the South and Southeast of the domain are essentially removed.

Each model is evaluated against the gridded observations over the flooding event period, November 3 12Z–November 7 12Z, 1994, and during the entire verification dataset from October 1 12Z to November 29 12Z, 1994, in terms of scatter diagrams (Fig. 5). The *R*-squared coefficient, RMSE, and bias values are also reported. The WRF model tends to overpredict significantly when compared with gridded observation points which were greater

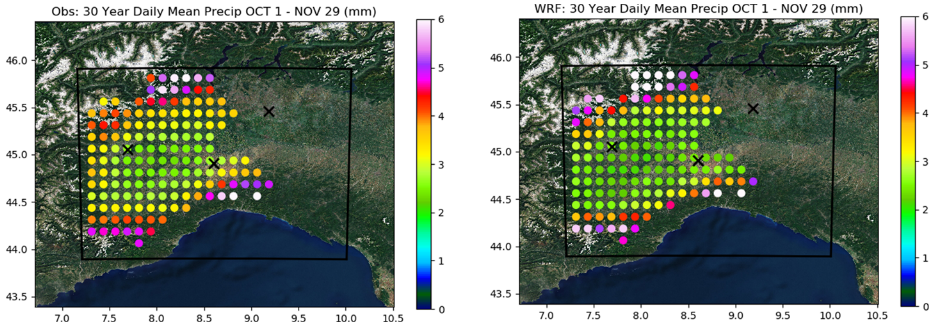
Obs: 30 Year Daily Mean Precip OCT 1 - NOV 29 (mm)
WRF: 30 Year Daily Mean Precip OCT 1 - NOV 29 (mm)
46.0
45.5
45.0
44.5
44.0
43.5
7.0
7.5
8.0
8.5
9.0
9.5
10.0
10.5
0
1
2
3
4
5
6

**Fig. 2** Thirty-year daily precipitation mean for October 1 12Z–November 29 12Z for gridded observations (left) and WRF (right). *X*s denote the cities of Torino, Alessandria, and Milano from left to right

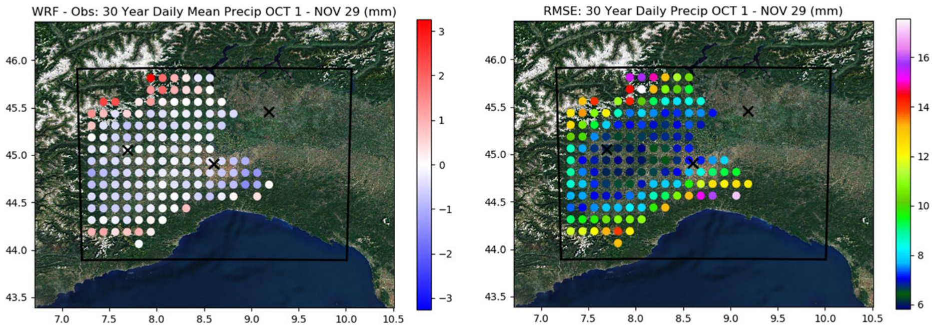
WRF - Obs: 30 Year Daily Mean Precip OCT 1 - NOV 29 (mm)
RMSE: 30 Year Daily Precip OCT 1 - NOV 29 (mm)
46.0
45.5
45.0
44.5
44.0
43.5
7.0
7.5
8.0
8.5
9.0
9.5
10.0
10.5
3
2
1
0
−1
−2
−3
16
14
12
10
8
6

**Fig. 3** Bias (left) and root mean squared error (right) of WRF against observations of the daily precipitation computed over a 30-year period from October 1 12Z to November 29 12Z. Xs denote the cities of Torino, Alessandria, and Milano from left to right

than 100 mm. There is an overpredictive bias during the 2-month simulation period and is more pronounced during the flooding event. When comparing the observations to the model estimates resulting from the AnEn, the tendency to significantly overpredict when the gridded observations display values greater than 100 mm is largely diminished comparatively. Especially during the flooding event, this reduces the bias considerably. The CNN technique further removes overpredictions during all types of rainfall events so much that a negative bias emerges. Precipitation events around 75 mm and greater are increasingly underpredicted the larger the precipitation total. Overall, the WRF model displays the highest RMSE and an overprediction bias. AnEn reduces the model bias the most (slightly positive). CNN shows the lowest RMSE, the highest *R*-squared value, but still tends to underpredict especially at higher precipitation totals.

## 6 Conclusion

In this study, we wanted to explore the potential of two different post-processing techniques to improve precipitation estimates from an NWP model such as WRF during the 1994 flood event over the Italian Piedmont region. To this purpose, a 30-year dataset for the months of October and November for the period 1984–2013 has been built using WRF 24-h accumulated precipitation (AP) amounts over Piedmont at a resolution of 3 km. WRF simulations have been driven by NOAA CFSR reanalysis fields used as initial and boundary conditions. The dataset also includes gridded AP data generated through optimal interpolation techniques from

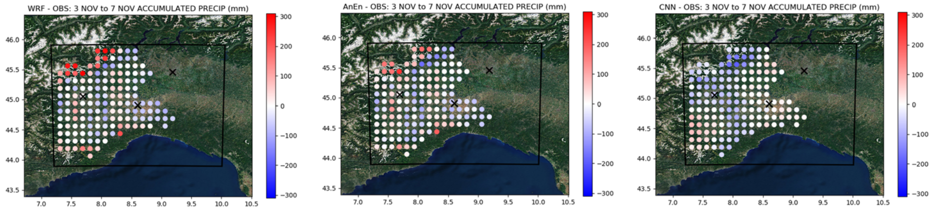
WRF - OBS: 3 NOV to 7 NOV ACCUMULATED PRECIP (mm)
AnEn - OBS: 3 NOV to 7 NOV ACCUMULATED PRECIP (mm)
CNN - OBS: 3 NOV to 7 NOV ACCUMULATED PRECIP (mm)
46.0
45.5
45.0
44.5
44.0
43.5
7.0
7.5
8.0
8.5
9.0
9.5
10.0
10.5
300
200
100
0
−100
−200
−300

**Fig. 4** Accumulated precipitation difference from observations from November 3 12Z to November 7 12Z, 1994 for WRF (left), analog ensemble (center), and convolutional neural network (right). Xs denote the cities of Torino, Alessandria, and Milano from left to right

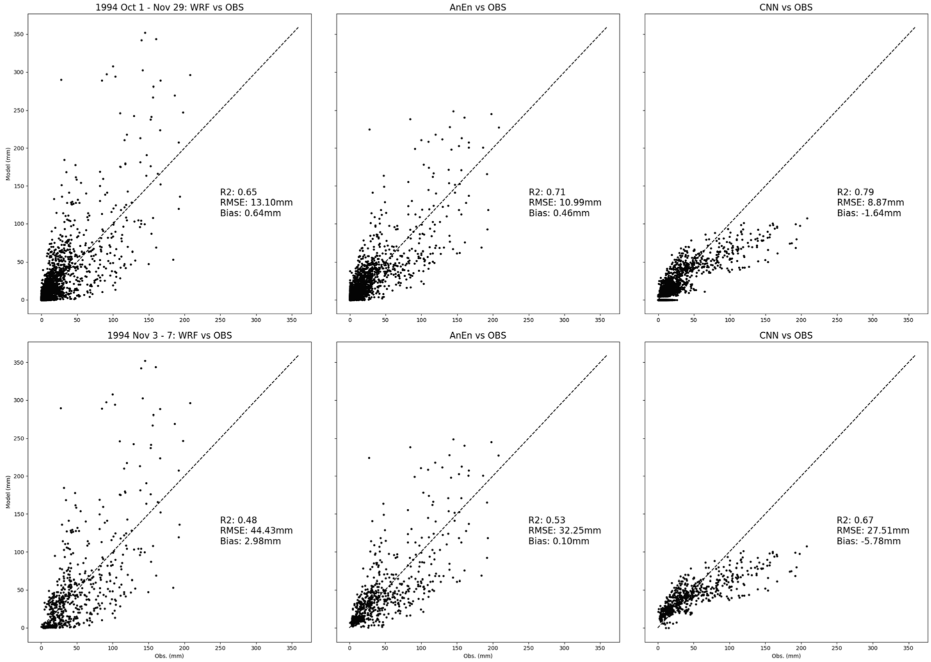

**Fig. 5** Scatter diagrams of simulated WRF daily precipitation totals against observations for WRF (left), AnEn (center), CNN (right) from October 1 12Z to November 29 12Z, 1994 (top), and November 3 12Z–7 12Z, 1994 (bottom). *R*-squared value, RMSE, and bias are also reported

the observed values of the Piedmont station network by the Forecasting Systems Department of ARPA Piemonte and referred to as the NWIOI dataset (Ronchi et al. 2008).

Climatologically, the WRF model has a tendency to underpredict accumulated daily precipitation in the interior Po valley of the Piedmont region of Northwestern Italy while more significantly overpredicting in the extreme Northern portion of the region in the Italian Alps. Specifically, the focus of this paper is the performance around a significant flooding event that occurred from November 3 12Z to 7 12Z, 1994, which had displayed unprecedented precipitation rates in the region, largely found in the complex topography surrounding the interior valley. Two different post-processing techniques were applied to the flood case to improve the precipitation results. The point-based AnEn and the grid-based CNN were able to reduce these large overprediction differences between the Piedmont rainfall observation network and the WRF model. Both techniques significantly improve the root mean squared error (RMSE) and the correlation of the precipitation fields with respect to WRF. However, the AnEn retained the properties of the exceptional nature of the event while the CNN introduces a strong underprediction bias conditional to observations, with an increasing underprediction for larger observed precipitation values. In fact, the CNN technique reduced the WRF overpredictions and smoothed out the local precipitation minima and maxima compared with the results of the AnEn and WRF.

In particular, the AnEn reduced the large overpredictions found locally in WRF in the extreme Northern portion of the Piedmont region which was surrounded by strong underpredictions, while the CNN seemed to spread the surrounding underpredictions possibly due to the use of spatial correlations among errors in this area. Even though this study has used a reanalysis precipitation dataset obtained through a dynamic downscaling with an NWP, we expect similar results in terms of relative performances between the AnEn (point-based) and CNN (grid-based) model output statistics techniques if these techniques were applied to forecast precipitation data. In that case, a dataset with NWP forecast data would be needed instead of a reanalysis dataset for the training. Hence, this study demonstrates the potential of these post-processing techniques to improve NWP precipitation estimates in case of a future flood event in the Piedmont region.

In terms of computational resources, both the AnEn and CNN techniques can run in less than a minute on a common personal computer to generate a single 0–24-h AP estimate. The computational burden is in the training process (the weight optimization in the AnEn's case) for both techniques which can require 2–3 h to be completed. The CNN has been trained on a single NVIDIA Tesla V100 GPU and training speeds could be vastly sped up by distributing to multiple GPUs. However, in an operational setup, the training can be carried out offline and not necessarily on a daily basis.

**Acknowledgments** The authors acknowledge the National Ground Intelligence Center (NGIC) of the US Army for supporting the development of the Global Climatological Analysis Tool (GCAT).

**Funding** The National Center for Atmospheric Research is sponsored by the National Science Foundation.

## References

Abadi M, Barham P, Chen J, et al (2016) TensorFlow: a system for large-scale machine learning. 12th USENIX Symposium on Operating Systems Design and Implementation. https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi

Alessandrini S, Delle Monache L, Sperati S, Nissen JN (2015) A novel application of an analog ensemble for short-term wind power forecasting. Renew Energy 76:768–781. https://doi.org/10.1016/j.renene.2014.11.061

Alessandrini S, Vandenberghe F, Hacker JP (2017) Definition of typical-day dispersion patterns as a consequence of a hazardous release. Int J Environ Pollut 62(2–4):305–318. https://doi.org/10.1504/IJEP.2017.089416

Alessandrini S, Delle Monache L, Rozoff CM, Lewis WE (2018) Probabilistic prediction of tropical cyclone intensity with an analog ensemble. Mon Weather Rev 146(6):1723–1744. https://doi.org/10.1175/MWR-D-17-0314.1

Alessandrini S, Sperati S, Delle Monache L (2019) Improving the analog ensemble wind speed forecasts for rare events. Mon Weather Rev 147(7):2677–2692. https://doi.org/10.1175/mwr-d-19-0006.1

Arpa (2010) Arpa Piedmont Forecasting System. http://www.arpa.piemonte.it/rischinaturali/tematismi/clima/confronti-storici/dati/. Accessed 15 Sept 2019

Binder P (1996) MAP — Mesoscale Alpine pro-stability to slantwise moist convection. J Atmos Sci 44:1559–1573

Buzzi A, Tartaglione N, Malguzzi P (1998) Numerical simulations of the 1994 piedmont flood: role of orography and moist processes. Mon Weather Rev 126(9):2369–2383. https://doi.org/10.1175/15200493(1998)126<2369:NSOTPF>2.0.CO;2

Cassardo C, Loglisci N, Gandini D, Qian MW, Niu GY, Ramieri P, Pelosini R, Longhetto A (2002) The flood of November 1994 in Piedmont, Italy: a quantitative analysis and simulation. Hydrol Process 16(6):1275–1299. https://doi.org/10.1002/hyp.1062

Chapman WE, Subramanian AC, Delle Monache L, Xie SP, Ralph FM (2019) Improving atmospheric river forecasts with machine learning. Geophys Res Lett 46(17–18):10627–10635. https://doi.org/10.1029/2019GL083662

Chollet F (2015) Keras: the python deep learning library ascl: 1806.022

Delle Monache L, Eckel FA, Rife DL, Nagarajan B, Searight K (2013) Probabilistic weather prediction with an analog ensemble. J Ametsoc Org 141(10):3498–3516. https://doi.org/10.1175/MWR-D-12-00281.1

Delle Monache L, Alessandrini S, Djalalova I, Wilczak J, Knievel JC, Kumar R (2020) Improving air quality predictions over the United States with an analog ensemble. Weather Forecast 35:2145–2162. https://doi.org/10.1175/WAF-D-19-0148.1

Doswell CA, Ramis C, Romero R, Alonso S (1998) A diagnostic study of three heavy precipitation episodes in the Western Mediterranean region. Weather Forecast 13(1):102–124. https://doi.org/10.1175/1520-0434(1998)013<0102:ADSOTH>2.0.CO;2

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

Ferretti R, Low-Nam S, Rotunno R (2000) Numerical simulations of the Piedmont flood of 4-6 November 1994. Tellus A 52(2):162–180. https://doi.org/10.1034/j.1600-0870.2000.00992.x

Gagne Ii DJ, Haupt SE, Nychka DW, Thompson G (2019) Interpretable deep learning for spatial analysis of severe hailstorms. Journals Ametsoc Org 147(8):2827–2845. https://doi.org/10.1175/MWR-D-18-0316.1

Hamill TM, Whitaker JS (2006) Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. Mon Weather Rev 134(11):3209–3229. https://doi.org/10.1175/MWR3237.1

Hsu K, Sorooshian S, Pan B, Aghakouchak A (2019) Improving precipitation estimation using convolutional neural network. Water Resour Res 55(3):2301–2321. https://doi.org/10.1029/2018WR024090

Junk C, Delle Monache L, Alessandrini S (2015) Analog-based ensemble model output statistics. J Ametsoc Org 143(7):2909–2917. https://doi.org/10.1175/MWR-D-15-0095.1

Keller JD, Delle Monache L, Alessandrini S (2017) Statistical downscaling of a high-resolution precipitation reanalysis using the analog ensemble method. Journals. Ametsoc. Org. 56(7):2081–2095. https://doi.org/10.1175/JAMC-D-16-0380.1

Kim S, Hong S, Joh M, Song S (2017) DeepRain: ConvLSTM network for precipitation prediction using multichannel radar data. 7th International Climate Informatics Workshop Sept. 20-22, 2017. arXiv preprint arXiv:1711.02316

Kingma D, Ba J (2015). Adam: a method for stochastic optimization. 3rd International Conference for Learning Representations 2015. http://arxiv.org/abs/1412.6980

Ladwig W (2017) Wrf-python (1.3.2). UCAR/NCAR. https://www.wrf-python.readthedocs.io/en/latest/

Lauret P, Diagne HM, David M, Diagne M (2014) A neural network post-processing approach to improving NWP solar radiation forecasts. Energy Procedia 57:1044–1052. https://doi.org/10.1016/j.egypro.2014.10.089ï

NCAR (2020) GCAT - system overview. https://ral.ucar.edu/solutions/products/global-climatology-analysis-tool-gcat. Accessed 15 Sept 2019

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

Powers JG, Klemp JB, Skamarock WC, Davis C, Dudhia J, Gill D, Coen J, Gochis DJ, Ahmadov R, Peckham S (2017) The weather researching and forecasting model: overview, system efforts, and future directions. Journals Ametsoc Org 98(8):1717–1737. https://doi.org/10.1175/BAMS-D-15-00308.1

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

Romero R, Ramis C, Alonso S, Doswell C, Stensrud D (1998) Mesoscale model simulations of three heavy precipitation events in the western Mediterranean region. Mon Weather Rev 126:1859–1881. https://doi.org/10.1175/1520-0493(1998)126%3C1859:MMSOTH%3E2.0.CO%3B2

Ronchi C, Luigi CD, Ciccarelli N, Loglisci N (2008) Development of a daily gridded climatological air temperature dataset based on a optimal interpolation of ERA-40 reanalysis downscaling and a local high resolution thermometers network. 8th EMS Annual Meeting and 7th European Conference on Applied Climatology Sept. 29 - Oct. 3 2008

Rotunno R (2001) Mechanisms of intense Alpine rainfall. J Atmos Sci 58(13):1732–1749. https://doi.org/10.1175/1520-0469(2001)058<1732:MOIAR>2.0.CO;2

Scher S (2018) Toward data-driven weather and climate forecasting: approximating a simple general circulation model with deep learning. Geophys Res Lett 45(22):12,616–12,622. https://doi.org/10.1029/2018GL080704

Skamarock W, Klemp J, Dudhia J, Gill D (2008) A description of the advanced research WRF version 3. NCAR technical note-475+ STR. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.484.3656

Sperati S, Alessandrini S, Delle Monache L (2017) Gridded probabilistic weather forecasts with an analog ensemble. Q J R Meteorol Soc 143(708):2874–2885. https://doi.org/10.1002/qj.3137

Toms BA, Kashinath K, Yang D (2019) Deep learning for scientific inference from geophysical data: the Madden-Julian oscillation as a test case. arXiv preprint arXiv:1902.04621

Weyn JA, Durran DR, Caruana R (2019) Can machines learn to predict weather? Using deep learning to predict gridded 500-hPa geopotential height from historical weather data. J Adv Model Earth Syst 11(8):2680–2693. https://doi.org/10.1029/2019MS001705