Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 10, No. 2, 2024

Benchmark Dataset Applied to Quantitative Precipitation Estimation and Forecasting of Qinghai

Yanping Li, Qingjun Yang *, Zhiyun Lai, Yu Qian

Qinghai Meteorological Information Center, Xining, Qinghai, China
* **Corresponding author**: Qingjun Yang (Email: qhyqj@126.com)

Abstract: We collected short-duration heavy precipitation events around weather radar since 2020 and get radar base data of these events. Then, we extracted radar products like Composite Reflectivity (CR), echo top height (ET) and Vertically Integrated Liquid Water (VIL) from radar base data after quality control as the input of dataset, and merge the surface precipitation of 6 minutes with background field using multi-grid variation to generate precipitation label by 6 minutes. Thus, we constructed two datasets artificial intelligence application QpefQH for short-duration heavy precipitation. QpefQH dataset has a scale of more than 15,000 images available for radar echo extrapolation and quantitative precipitation estimation tasks based on deep learning. QpefQH dataset also includes multi-source data such as satellite, numerical model, sounding, lightning, land surface temperature, pressure, wind and other observation data. QpefQH can be shared as a benchmark dataset for artificial intelligence research on short-duration heavy precipitation in Qinghai Province.

Keywords: Short-duration Heavy Precipitation; Quantitative Precipitation Estimation; Radar Echo Extrapolation.

1. Introduction

Short-duration heavy precipitation is a kind of strong convective weather, and it has a characteristic of suddenness, locality and destructiveness. It is easy to cause meteorological disasters, and then lead to secondary disasters such as debris flow and landslide. Short-duration heavy precipitation has occurred in Qinghai area frequently. Since Qinghai-Tibet Plateau trends to warm and humid climate, the frequency of short-duration heavy precipitation has increased. Short-duration heavy precipitation is one of the meteorological disaster that needs to be prevented. On August 18, 2022, a flash flood caused by short-duration heavy precipitation in Datong County leaded to casualties and direct economic losses of 690 million. Since September 3 in 2024, a heavy rain event happened in Qinghai province. Seven meteorological stations have seen a precipitation of more than 100 mm. This weather event was the strongest rainstorm since any meteorological records occurred. Short-duration heavy precipitation usually has a small scale and develops rapidly. It has always been a difficult problem in numerical model weather forecast. The rapid development of deep learning theory has brought a new way for this problem. Compared with the numerical model forecast, deep learning has shown comparable performance and faster inference speed than numerical model in precipitation forecast.

Doppler weather radar can provide cloud and rain observation with a spatial resolution of 1 km and time resolution of 6 minutes in all weather conditions. It can explore 9 layers in vertical direction, and observe with a distance reach up to 230 km within the effective observation. It has a high frequency of observation and wide coverage through a weather event. It can be an effectively supplement observation to the blind area of surface observation in Qinghai area. Thus, the weather radar has become the most important observation method for short-duration heavy precipitation. However, the training of deep learning network relies on long-sequence and high-quality training datasets, and there is no available open source short-duration heavy precipitation dataset for artificial intelligence application in Qinghai. In this paper, we constructed a training dataset QpefQH for artificial intelligence application on short-duration heavy precipitation in Qinghai, aiming at monitoring and forecasting short-duration heavy precipitation in Qinghai. The dataset consists of two subsets, one for radar echo extrapolation and one for quantitative precipitation estimation. The dataset lays a data foundation for research on short-duration heavy precipitation with artificial intelligence method in Qinghai.

2. Related Research

Many researchers have tried to apply deep learning method on precipitation forecasting. MetNet[1] used satellites, radars, and other data as inputs, and can predicted the precipitation probability of the next 8 hours with a spatial resolution of 1km and time resolution of 2 minutes. MetNet has higher accuracy and faster inference speed comparing with the physical model HREF. MetNet-2[3] maintained the spatial resolution of MetNet and extended the forecast time to 12 hours. MetNet-2 outperforms HREF in both high and low precipitation amounts. DGMR [4] network used a generative adversarial network structure, and provided precipitation probabilities of the next 90 minutes in 1 second. DGMR balances the intensity and scope of precipitation forecasts, and captures the strength and structure of circulation better. In recent years, with the enhancement of floating-point computing ability, deep learning networks applied to weather forecasting trended to large models. Pangu[5] designed a 3D transformer structure to generate global forecast with a resolution of 0.25° ×0.25°. The forecast results of Pangu are superior to the numerical model IFS and has a faster inference speed. Fuxi[6] can generate global forecasts of the next 15 days every 6 hours at a resolution of 0.25° ×0.25°. The forecast results of Fuxi have a performance comparable to the EM model. Fengwu[7] can generate global forecast results of the next 10 days in 30 seconds at a resolution of 10KM. The forecast results of Fengwu is superior to GraphCast[8].

Several datasets have been conducted to satisfy the need of artificial intelligence application in meteorology. ERA5[9] is commonly used as data set in training large meteorological models. ERA5 can provide hourly data of land surface and upper-air parameters from 1940 to the present. ERA5 has a horizontal resolution of about 31 kilometers and a total number of 137 levels. LiuNa[10] collected strong convective cases of China from 2012 to 2019, and constructed a basic training dataset SCWDS for artificial intelligence applications in strong convective. SCWDS dataset included 184,865 cases of 5 types of strong convective weather, such as thunderstorm, thunderstorm gale, short-duration heavy precipitation, hail, and tornadoes. Xiong[11] used strong precipitation weather events of six provinces in central and southern China from 2016 to 2018 to construct a minute-level quantitative precipitation estimation dataset QpefBD. QpefBD included 3,185 strong precipitation cases and a total number of over 230,000 images. Xie[12] constructed a dataset for identifying squall lines on radar echo. The dataset included 12 squall line in North and Southeast China from 2019 to 2021, and had a total number of about 50,000 images. Tian[12] constructed a dataset for the identifying gust fronts on radar echoes. The dataset included 1,422 gust fronts in North and Southeast China, and had a total number of over 27,000 images.

3. Overview of Research Area

Qinghai Province is located in the Northeast area of the Qinghai-Tibet Plateau with longitudes ranging from 89°35′ to 103°04′ and latitudes from 31°36′ to 39°19′. Qinghai has an average elevation of 4050 meters and a plateau continental climate feature. Qinghai Province has a total number of 1082 meteorological observation stations by the end of 2023 with observation frequency reaching minute level. Qinghai Province also has 5 Doppler radars, 7 sounding stations, and 33 lightning observation stations and new types of observational equipment such as X-band radars, microwave radiometers, wind profilers, and millimeter-wave cloud radars. There are also plenty of Fengyun and Himawari series of satellites observation data distributed by China Meteorological Administration, along with various model data such as CMA model and ECMWF model. The data used in this study all comes from Qinghai provincial and national meteorological big data cloud platforms.

4. Data and Processing

4.1. Data Collection

The Center Meteorological Observatory defines short-duration heavy precipitation as precipitation exceeding 20mm within 60 minutes. According to the meteorological department's ground observation specifications, the hourly precipitation defines as total precipitation between 01 minutes and 00 minutes of the next hour. We first extracted hourly precipitation exceeds 20mm within the radar observation window. Then we merge those hours of intervals no more than 2 hours into the same process. Finally, we take 2 hours before and after the occurrence of processes as the time window of this process. We define the approximately 320km×320km with the radar at the center as observation window. Besides, we search the time and location of the flood disaster caused by heavy rain from the disaster reporting system as a supplement to the processes extracted from the meteorological big data cloud platform.

4.2. Data Filtering

We further filtered all the weather events collected in section 4.1. We established several general filtering principles as follows. (1) We exclude processes with more than missing data of 3 volume scans or data format errors in the radar base data. (2) If main body of radar echo is not within the radar observation window or located near the boundary of the radar's effective observation range, such processes should be excluded. We applied different filtering principles considering the different use of two subsets. For the dataset used for radar echo extrapolation, the main consideration is focused on temporal characteristics, so processes that last no more than 3 hours are excluded. For the dataset used for quantitative precipitation estimation by radar, the main consideration is focused on the accuracy and completeness of the label. So, we excluded processes with low intensity such as processes with no more than 2 surface stations meeting the short-duration heavy rainfall standard. We also excluded processes if the maximum hourly precipitation during the process is slightly higher than 20mm. Besides, we excluded processes with few ground stations or unevenly distributed during the short-duration heavy precipitation event occurs.

4.3. Feature Selection

The dataset includes information from radar, satellite, numerical model, atmospheric reanalysis, surface observation, raindrop spectrum, real-time data, lightning observation, sounding data and other data. We identified main features of each data by literature research and expert discussions. Radar features include base data and CR, ET, VIL. Surface observation features include various instantaneous observation values and extreme values of temperature, pressure, humidity, and precipitation. Satellite data includes secondary products of FengYun-4A such as cloud phase, black body temperature and other product. The details of the various data features in the dataset have shown in table 1.

Table 1. Features in QpefQH

No.	Data	Frequency	Main Features
1	Radar	6min	Base data, Composite Reflectivity, Echo Top, Vertically Integrated Liquid
2	Surface	1 or 5min	Temperature, Pressure, Humidity, Wind and Precipitation.
3	Satellite	15min	Cloud Phase, Cloud Type, Quantitative Precipitation Estimation.
4	Numerical Model	12h	Convective Available Potential Energy, Total Precipitation, Temperature.
5	Reanalysis Data	1h	Convective Available Potential Energy, Total Precipitation, Temperature.
6	Raindrop Spectrum	When raining	Particle Diameter Velocity Spectrum.
7	Lightning	When lightning	Longitude, Latitude, Altitude, Intensity.
8	Sounding Data	12h	Altitude, Humidity, Wind Direction, Wind Speed, Temperature
9	Real-time Data	10min or 1h	Hourly Precipitation, Precipitation in 10min.

4.4. Data Quality Control

All the data used in the dataset have undergone strict

quality control. The quality control methods for surface observation data include boundary checking, missing value imputation, and consistency checking. Boundary checks examine values that exceed a reasonable range and sets them as missing value, such as hourly precipitation exceeding 100mm. Consistency checking include temporal consistency and spatial consistency. Temporal consistency examines if the observational value from the same station is consistent with the previous and next value. Spatial consistency examines if the observations at the same time is consistent with the values of nearby stations in space. Observation values that significantly deviate from consistency range are eliminated and set as missing value. Missing value imputation refers to using data from the neighbor time at the same station or from neighbor stations at the same time to interpolate the missing value according to consistency condition.

Quality control for radar base data includes attenuation correction, velocity dealiasing, clutter suppression, ground echo elimination, and missing value imputation. First, we corrected the radar base reflectivity attenuation by using the HB method. Then, we suppressed the clutter by using the correlation coefficient (CC) of the H and V components of the dual-polarization radar. Radar echo with a correlation coefficient lower than 80% is considered non-meteorological echoes and should be eliminated. Finally, we extracted CR, ET and VIL from radar base data after quality control. We collected other radar products such as HCL from PUP product uploaded by each radar station. We corrected the radar radial velocity by using a velocity dealiasing algorithm. Besides, we interpolate the missing one volume scan by using the previous and next scans. The process of performing quality control on radar data has shown in Figure 1.

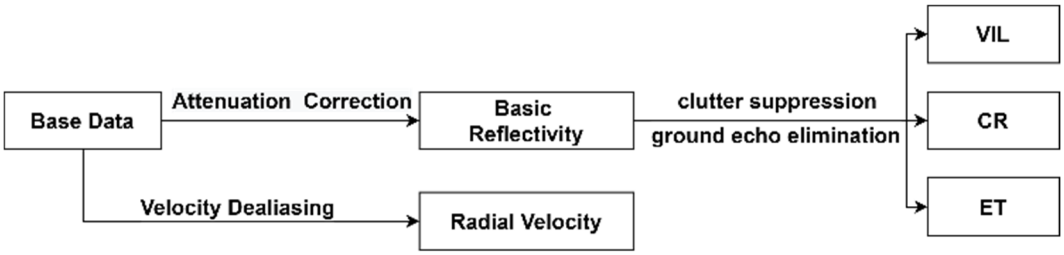

Figure 1. The quality control process of radar data.

4.5. Generating Labels

The radar echo extrapolation dataset uses previous N volume scans as input and the next volume scan as the label. It is sufficient to perform quality control on the radar echoes. The quantitative precipitation estimation uses radar echo as input and precipitation intensity and areas of surface as label. We generated this label by using a multigrid variation method. The multigrid variation STMAS [13] is a new generation of fusion systems based on multigrid variation developed by XieYuanfu from NOAA Earth System Research Laboratory. This system combines the advantages of multiscale analysis and statistical analysis, and uses variation methods to simulate traditional objective analysis. STMAS applies appropriate balance and dynamic constraints at different analysis scales. It overcomes the uncertainties errors of the background field covariance errors matrix.

Due to poor communication signals, nearly half even more of stations in Qinghai operate at a 5-minute frequency. To make full rational use of these observations, we analyzed the minute precipitation and precipitation minutes last in each hour during short-duration heavy precipitation events in Qinghai Province. We found that short-duration heavy precipitation in Qinghai Province mostly shown as small minute precipitation lasts for a longer period of time. Therefore, the precipitation by 6 minutes on 5-minute observations could been estimated from the neighbor 5-minute periods according to the time proportion in each 5 minutes. Otherwise, we summed up minute precipitation for stations with 1-minute observations to get the precipitation by every 6 minutes.

In the beginning of each short-duration heavy precipitation events, we use hourly real-time precipitation as background field, then merge with 6-minute precipitation of all stations by using STMAS method. Then we use the previous 6-minutes precipitation label as the background field, and merge with 6-minute precipitation from surface stations by using STMAS method to get next precipitation label. The precipitation intensity and precipitation areas labels throughout the entire process could be achieved by repeating this cycle. The process of generating precipitation labels has been shown in Figure 2.

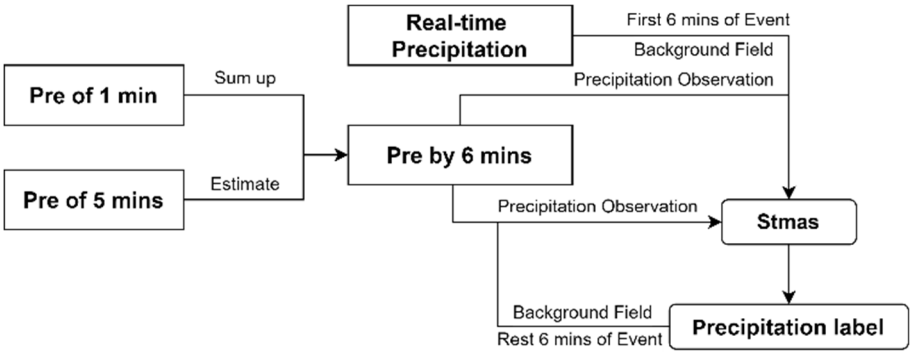

Figure 2. The process of generating label.

4.6. Data Normalization

We performed post-processing measures on various data in the dataset before sharing the dataset. The measures include standardization of elevation angles, filenames, and formats. The completion time and elevation angles used in each volume scan not always the same because of different observation modes. We standardized elevation angles of base reflectivity in the dataset to the following six angles: 0.5°, 1.4°, 2.4°, 3.4°, 4.5°, and 6.0°. The observation time interval between two volume scan is standardized to exactly 6 minutes. Finally, we save the data in npz format to preserve the decimal values of observation.

5. Dataset Overview

The QpefQH dataset has included all the short-duration heavy events within each Doppler weather radar's effective observation radius in Qinghai since 2020. The dataset was divided into two subsets: radar echo extrapolation dataset and quantitative precipitation estimation dataset. The dataset can be directly used for training deep learning networks. The input and output of radar echo extrapolation dataset are both radar echo data. The input of the quantitative precipitation estimation dataset consists of several radar product such as CR, ET and VIL, while label of the dataset is surface precipitation intensity and area markers. The rough labels only include the surface precipitation area, and the fine labels includes surface precipitation intensity. The scale of the dataset is shown in table 2.

Table 2. Scale of QpefQH

Dataset	Event Counts	Echo Counts	Label
Radar Echo Extrapolation	164	15788	Echo of next volume scan.
Quantitative Precipitation Estimation	249	16303	Precipitation intensity and precipitation area.

6. Summary and Outlook

We collected short-duration heavy precipitation events in Qinghai from 2020 to 2024, and then performed strict quality control on both radar and surface observation data to conduct a dataset for artificial intelligence application. The dataset contains multi-source data such as radar, satellite, numerical model, atmospheric reanalysis data and other data. This dataset can support deep learning applications for radar echo extrapolation and quantitative precipitation estimation. In the future we will further optimize of the dataset based on its performance in training radar echo extrapolation and quantitative precipitation estimation networks. We will continue to add new short-duration heavy precipitation events of future year. We hope this dataset could play a role in artificial intelligence research in Qinghai meteorology field and make a contribution to the construction of intelligent meteorology.

Acknowledgments

This research was supported by the National Meteorological Information Center's open competition mechanism project "Construction of Artificial Intelligence Application Training Dataset for Severe Convective Weather on Qinghai-Tibet Plateau" (09), Qinghai provincial meteorological bureau's key project "Construction of Artificial Intelligence Application Training Dataset for Short-duration Heavy Precipitation in Qinghai" (QXZD2024-03) and Global Atmospheric Background and Qinghai-Tibet Plateau Big Data Application Center Science and Technology Innovation Platform (2023-SF-10).

References

[1] The Ministry of Emergency Management has released the "Top Ten Natural Disasters in China for the Year 2022" [J]. Safety and Health, 2023, (04): 53-54.

[2] Sønderby, C.K., Espeholt, L., Heek, J., Dehghani, M., Oliver, A., Salimans, T., Agrawal, S., Hickey, J., & Kalchbrenner, N. (2020). MetNet: A Neural Weather Model for Precipitation Forecasting. ArXiv, abs/2003.12140.

[3] Espeholt, L., Agrawal, S., Sønderby, C.K., Kumar, M., Heek, J., Bromberg, C., Gazen, C., Hickey, J., Bell, A., & Kalchbrenner, N. (2021). Skillful Twelve Hour Precipitation Forecasts using Large Context Neural Networks. ArXiv, abs/2111.07470.

[4] Ravuri, S., Lenc, K., Willson, M. et al. Skilful Precipitation Nowcasting Using Deep Generative Models of Radar. Nature 597, 672–677 (2021).

[5] Bi, K., Xie, L., Zhang, H. et al. Accurate medium-range global weather forecasting with 3D neural networks. Nature 619, 533–538 (2023).

[6] Chen, L., Zhong, X., Zhang, F. et al. FuXi: a cascade machine learning forecasting system for 15-day global weather forecast. npj Clim Atmos Sci 6, 190 (2023).

[7] Chen K, Han T, Gong J, et al. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead[J]. arXiv preprint arXiv:2304.02948, 2023.

[8] Lam R, Sanchez-Gonzalez A, Willson M, et al. GraphCast: Learning skillful medium-range global weather forecasting[J]. arXiv preprint arXiv:2212.12794, 2022.

[9] Hersbach H, Bell B, Berrisford P, et al. The ERA5 global reanalysis[J]. Quarterly Journal of the Royal Meteorological Society, 2020, 146(730): 1999-2049.

[10] QpefBD: A Benchmark Dataset Applied to Machine Learning for Minute-Scale Quantitative Precipitation Estimation and Forecasting[J]. Journal of Meteorological Research, 36(1):93-106.

[11] Liu Na, Xiong Anyuan, Zhang Qiang, et al. Construction of Basic Dataset for Artificial Intelligence Application Training on Severe Convective Weather.[J] Journal of Applied Meteorological Science, 2021, 32(5): 530-541.

[12] Xie P, Hu Z, Yuan S, et al. RADAR Echo Recognition of Squall Line Based on Deep Learning[J]. Remote Sensing, 2023, 15(19): 4726.

[13] Tian H, Hu Z, Wang F, et al. Radar Echo Recognition of Gust Front Based on Deep Learning[J]. Remote Sensing, 2024, 16(3): 439.