
IowaRain: A Statewide Rain Event Dataset Based on Weather Radars and Quantitative Precipitation Estimation

Muhammed Sit¹ Bong-Chul Seo¹ Ibrahim Demir¹

Abstract

Effective environmental planning and management to address climate change could be achieved through extensive environmental modeling with machine learning and conventional physical models. In order to develop and improve these models, practitioners and researchers need comprehensive benchmark datasets that are prepared and processed with environmental expertise that they can rely on. This study presents an extensive dataset of rainfall events for the state of Iowa (2016-2019) acquired from the National Weather Service Next Generation Weather Radar (NEXRAD) system and processed by a quantitative precipitation estimation system. The dataset presented in this study could be used for better disaster monitoring, response and recovery by paving the way for both predictive and prescriptive modeling.

1. Introduction

Many problems in the environmental domain can be tackled with data-driven approaches. However, as with any data-driven problem, the lack of available data limits the development and applicability of environmental modeling efforts. Furthermore, environmental modeling, specifically probability-based environmental modeling, in contrast with its dependency on data, suffers from data scarcity (Sit et al., 2020). The data scarcity problem in the field is being iterated in many studies (Ebert-Uphoff et al., 2017; Rolnick et al., 2019) while the cardinality of publicly available pre-processed datasets remains to be low.

1.1. Related Work

Studies rectify data scarcity one step at a time. CAMELS being one of the oldest and widely used datasets in rainfall-

runoff prediction tasks, provides a go-to mark for practitioners working on flood forecasting for contiguous United States (Newman et al., 2015). CAMELS provides streamflow measurements from the nation-wide United States Geological Survey (USGS) river sensor network as well as essential data points that are typically used in rainfall-runoff modeling. Although CAMELS is not prepared with statistical predictive modeling in mind to support physical modeling, another study, FlowDB (Godfried et al., 2020) presents two extensive datasets with a machine learning focus. The first dataset, following in CAMEL’s footsteps, focuses on streamflow levels along with precipitation measurements for flood forecasting purposes. FlowDB also goes beyond streamflow with the second dataset on flash flood events. The second dataset includes many aspects of historic flood events, including river flow, precipitation measurements, property damage estimations, injury data, and event locations. While both CAMELS and FlowDB involve precipitation measurements, their main focus falls into the flood forecasting domain rather than pure rainfall and precipitation context.

One benchmark dataset focusing on rainfall is WeatherBench (Rasp et al., 2020) which presents a dataset with a focus on medium-range weather forecasting pre-processed for deep learning applications. WeatherBench uses measurements from ERA5 numerical reanalysis dataset (Hersbach et al., 2020). ERA5 combines a forecast model with the available observations and provides estimations of the atmospheric state. An alternative dataset using ERA5 is RainBench (de Witt et al., 2020). In addition to numerical estimations from ERA5, RainBench also incorporates simulated satellite data as imagery and global precipitation estimates from IMERG (Huffman et al., 2015) to support easier data-driven weather forecasting.

Even though the number of datasets published directly or remotely related to precipitation has been gaining momentum for the last few years, the need for alternative environmental monitoring datasets in order to ensure better climate change modeling and planning is critical (Rolnick et al., 2019; Sit et al., 2020). This study presents a rain event dataset consisting of precipitation data between 2016 and 2019 from the state of Iowa based on the NEXRAD radar network

¹IIHR—Hydroscience Engineering, The University of Iowa, Iowa City, Iowa, USA. Correspondence to: Muhammed Sit <muhammed-sit@uiowa.edu>.

from seven radars positioned in and around the state. The main goal of this dataset is to support climate change modeling for effective disaster planning, response, and recovery. While the dataset presented here does not necessarily define a problem, it invites machine learning researchers to better study precipitation and rainfall by providing the dataset and various sample problems.

This paper is structured as follows; the introduction section briefly presents the problem that this paper tackles and covers previous works. The following section (Dataset) discusses the specifics of the dataset by describing how the raw data was acquired and processed. Then it shares descriptive statistics representing the dataset to form an understanding while listing potential environmental domain problems for that the dataset can be used. Lastly, in the conclusion section, we share the final remarks.

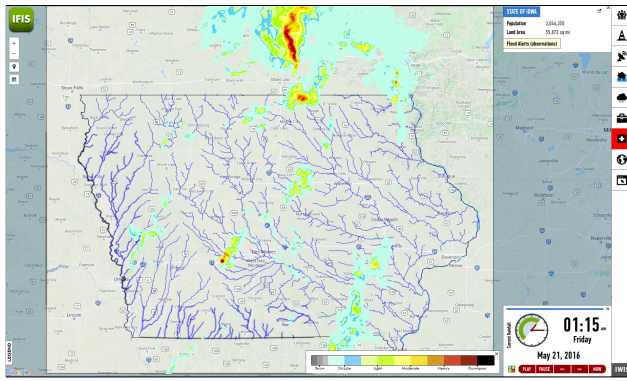


Figure 1. A 2D rain rate map visualized for Iowa domain.

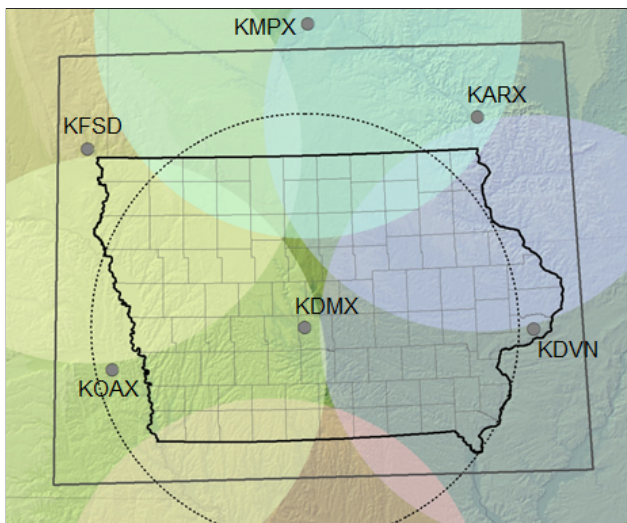


Figure 2. WSR-88D network radar coverages over Iowa domain.

2. Dataset

The IowaRain dataset comprises rainfall events from 2016 through the end of 2019 based on Iowa Flood Center's (IFC) Quantitative Precipitation Estimation (QPE) system. The IFC QPE system, for the most part, relies on the National Weather Service's Weather Surveillance Radar-1988 Doppler (WSR-88D) network, which is operational at over 160 locations in the United States. WSR-88Ds complete a scan every 4-5 minutes for approximately 80 nautical miles (nm) range and produce data representing precipitation within their respective domains. The generated QPE products are fed to the Rainfall Event Detection system to determine and separate rainfall events.

The IFC QPE system (Seo et al., 2019) uses raw data acquired in real-time from seven WSR-88D radars (KARX in La Crosse, Wisconsin; KDMX in Des Moines, Iowa; KDVN in Davenport, Iowa; KEAX in Kansas City, Missouri; KFSO in Sioux Falls, South Dakota; KMPX in Minneapolis, Minnesota; and KOAX in Omaha, Nebraska). In addition, the system incorporates the specific attenuation algorithm (Seo et al., 2020) to ensure the efficient employment of WSR-88D's dual-polarization functions to improve QPE accuracy. To do so, it also determines melting layer locations by using temperature soundings from the numerical weather prediction models.

2.1. QPE System & Rainfall Event Detection

Using the WSR-88D network, the QPE system generates a composite rain rate map (Figure 1) covering the IFC domain that bounds the state of Iowa with buffer zones encircling the state (Figure 2), with 5 minutes of temporal and 500 meters of spatial resolutions, using a variety of processing algorithms. The unit of the product is millimeters/hour (mm/hr).

These QPE products are matrices, named snapshots, with the shape of 1057 x 1741 for each timestamp in [2016-2019] that have their minute component divisible by 5. Even though the QPE system produces maps for snow as well, the IowaRain dataset makes a distinction and only comprises rainfall events from April to October each year to avoid any potential derangement.

Data acquisition and processing for IowaRain are summarized in (Figure 3). The data manager module of the QPE system handles the acquisition of radar observables (reflectivity, radial velocity, spectrum, width, differential reflectivity, copolar correlation coefficient, and differential phase) as well as temperature and geo-potential height analysis for various pressure levels from numerical weather prediction models. After the data acquisition, the Data Manager ensures that the data is complete.

Once the data is ready, it goes into a series of process-

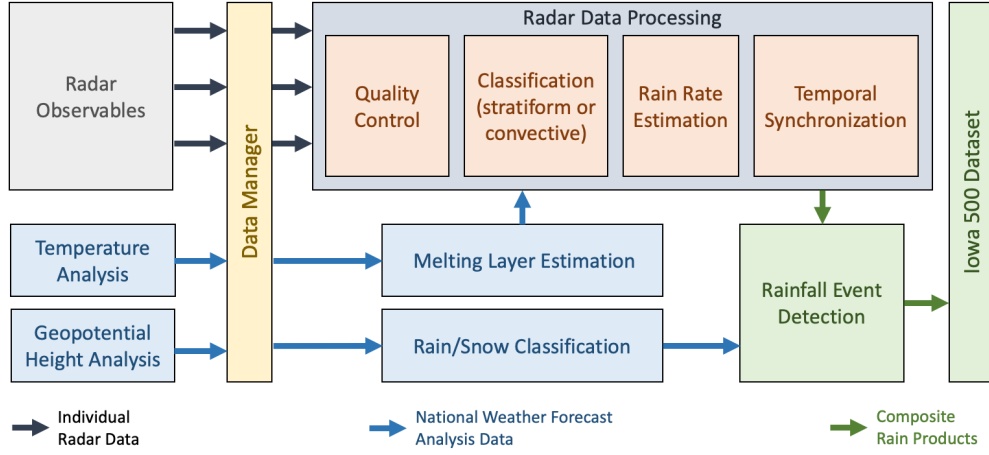


Figure 3. Overall summary of dataset acquisition and processing workflow.

ing sub-modules within the Data Processing subsystem for quality control, rain rate estimation, merging of individual radar data, and finally, rainfall product generation. In the quality control step, the system essentially eliminates non-meteorological data radars acquired, i.e., noise. After the data quality check, the QPE system generates temperature soundings for areas that radars cover by utilizing the numerical weather prediction analysis and calculates the altitude of the melting layer using 3D spherical coordinates. The system then runs a series of algorithms to classify individual radar scans into one of convective and stratiform using melting layer information.

The melting layer and classification of precipitation are then used by an estimator to generate 2D rain rate maps separate for each elevation. These individual elevation maps then get combined by a non-parametric kernel function to build the final 2D product for each radar scan. These 2D products built from individual radars need to be combined into one 2D product for the Iowa domain at the last step. Beyond the straightforward combination of various 2D maps into one, this step needs spatial and temporal synchronization as radars differ from each other with their spatial and temporal coverages. By combining 2D radar scans according to characteristics of the radar they originate from, the QPE system produces a composite 2D rain rate map every 5 minutes. The details regarding the QPE system and discussion about its accuracy can be reviewed further in (Seo & Krajewski, 2020). The composite rain rate maps are plain data files that need serialization for easier usage in data-driven tasks. To facilitate that, the Rainfall Detection System is used to read all the 2D plain rain rate maps into NumPy (Harris et al., 2020) arrays and then to be normalized by dividing them to 105, which records the presumably maximum observation in the Iowa domain. Since the matrices/snapshots the

QPE system produces have odd shapes, we resized them to (1088, 1760) by adding empty elements to enable more straightforward utilization of architectures that use iterative downsampling. This design choice could easily be reverted by removing extra rows and columns.

In order to create the final dataset, the serialized snapshots go through rainfall detection criteria. The criteria for a set of snapshots are defined as follows;

- All snapshots in the set must be consecutive
- The set must consist of at least ten snapshots
- Each snapshot must have at least 0.5 mm/hr precipitation over %10 of the domain.

The final version of the dataset consists of gzip files, one archive for each rain event. When any of the gzip archives is opened in Python with pickle, each file reads a triple that is structured as (*start_date, number_of_snapshots, list_of_snapshots*). The IowaRain dataset described in this section and sample code to open individual rainfall events can be found in the Git repository at <https://github.com/uihilab/IowaRain>.

Table 1. Number of events and average event duration by year.

YEAR	NUMBER OF EVENTS	AVG EVENT LENGTH
2016	64	7.09 HOURS
2017	67	6.55 HOURS
2018	76	8.04 HOURS
2019	81	7.09 HOURS

Table 2. Shortest and longest rainfall events by year.

YEAR	MIN EVENT LENGTH	MAX EVENT LENGTH
2016	50 MINUTES	29 HOURS
2017	50 MINUTES	29 HOURS, 15 MINUTES
2018	50 MINUTES	33 HOURS, 55 MINUTES
2019	50 MINUTES	17 HOURS, 40 MINUTES

2.2. Dataset Statistics

The IowaRain dataset comprises 1242 rainfall events with various durations. Since each year has different weather characteristics, there is no normal distribution of rainfall events year by year. While there is an increasing trend, the number of events each year is not drastically different (Table 1). The length of the longest rain event in 2019 is lower than the previous year by a large margin, while the minimum event length stays the same per the aforementioned event criteria (Table 2). A summary of events can be seen in Table 3, where we present average coverage of the domain and average rain rate in mm/hr by month and year.

Table 3. Average rainfall coverage and average precipitation of events by year and month.

YEAR	MONTH	AVG COVERAGE	AVG PRECIPITATION
2016	APR	16.81%	0.6370
	MAY	15.18%	0.6741
	JUN	12.90%	0.9355
	JUL	17.12%	1.3949
	AUG	13.56%	1.0103
	SEP	16.05%	1.0858
	OCT	17.25%	0.8573
2017	APR	14.01%	0.5657
	MAY	16.50%	0.7977
	JUN	13.00%	1.0379
	JUL	14.11%	0.9905
	AUG	14.16%	0.9263
	SEP	14.56%	0.6125
	OCT	14.32%	0.6791
2018	APR	14.60%	0.3699
	MAY	17.01%	0.9445
	JUN	14.94%	0.9249
	JUL	12.93%	0.8124
	AUG	14.04%	0.7316
	SEP	16.63%	0.9452
	OCT	13.80%	0.4725
2019	APR	15.26%	0.5559
	MAY	16.88%	0.7453
	JUN	14.47%	0.6307
	JUL	11.94%	0.8102
	AUG	15.34%	1.0646
	SEP	16.59%	1.0869
	OCT	17.93%	0.6226

Even though in data-driven environmental modeling, train/validation/test split is typically done by date like many other time-series focused problems, we do not necessarily put a constraint on that in this study. Since each rainfall event carries different characteristics, we leave the dataset splitting to other researchers' preferences.

2.3. Sample Tasks

It is needless to say that weather and rainfall forecasting typically are done by using more than just previous precipitation measurements. Nevertheless, IowaRain presents a unique challenge in rainfall forecasting, that is, forecasting rainfall events both short-term and long-term with limited data access. This problem is very much like the video frame prediction task, and since frame prediction is a well-studied problem, extensions of the state-of-the-art architectures for the video frame prediction problem could be explored with IowaRain.

Another potential task that might be traversed is about the dataset's nature. For example, as IowaRain is based on the WSR-88D network, it is prone to artifacts caused by the noise that is confusing radars. Those artifacts could be identified, and snapshots with artifacts could be labeled as an extension, and a data cleaning focused unsupervised approach could remove artifacts from snapshots.

One other outlook for IowaRain focuses on streamflow. Rainfall-runoff forecasting is one of the most studied problems in hydrosience with both physical-based and machine learning models. IowaRain enables practitioners and researchers to study streamflow by combining IowaRain with publicly available USGS sensor networks. By incorporating these datasets together, one could work on streamflow forecasting for both gauged and ungauged locations.

3. Conclusions

This paper presents a rainfall event dataset, namely [IowaRain](#), generated from weather radars and various other environmental analysis models. First, we briefly explained how the weather products from several radars covering areas in Iowa are acquired, processed, synchronized, and combined into 2D rain rate maps with temporal and spatial resolutions of 5 minutes and 500 meters, respectively. Subsequently, we described how individual rain events were determined and sampled how IowaRain could be helpful by outlining potential use cases as a starting point. Even though in its current form, the dataset covers only the state of Iowa and 2016-2019 as the timeframe, the same methodology described in this paper could be employed to extend the dataset for more comprehensive temporal and spatial coverage for the entire US.

References

- de Witt, C. S., Tong, C., Zantedeschi, V., De Martini, D., Kalaitzis, F., Chantry, M., Watson-Parris, D., and Bilinski, P. Rainbench: Towards global precipitation forecasting from satellite imagery. *arXiv preprint arXiv:2012.09670*, 2020.
- Ebert-Uphoff, I., Thompson, D. R., Demir, I., Gel, Y. R., Karpatne, A., Guereque, M., Kumar, V., Cabral-Cano, E., and Smyth, P. A vision for the development of benchmarks to bridge geoscience and data science. In *17th International Workshop on Climate Informatics*, 2017.
- Godfried, I., Mahajan, K., Wang, M., Li, K., and Tiwari, P. Flowdb a large scale precipitation, river, and flash flood dataset. *arXiv preprint arXiv:2012.11154*, 2020.
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730): 1999–2049, 2020.
- Huffman, G. J., Bolvin, D. T., Braithwaite, D., Hsu, K., Joyce, R., Xie, P., and Yoo, S.-H. Nasa global precipitation measurement (gpm) integrated multi-satellite retrievals for gpm (IMERG). *Algorithm Theoretical Basis Document (ATBD) Version*, 4:26, 2015.
- Newman, A., Clark, M., Sampson, K., Wood, A., Hay, L., Bock, A., Viger, R., Blodgett, D., Brekke, L., Arnold, J., et al. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous usa: data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, 19(1):209–223, 2015.
- Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., and Thuerey, N. Weatherbench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020.
- Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., et al. Tackling climate change with machine learning. *arXiv preprint arXiv:1906.05433*, 2019.
- Seo, B.-C. and Krajewski, W. F. Statewide real-time quantitative precipitation estimation using weather radar and nwp model analysis: Algorithm description and product evaluation. *Environmental Modelling & Software*, 132: 104791, 2020.
- Seo, B.-C., Keem, M., Hammond, R., Demir, I., and Krajewski, W. F. A pilot infrastructure for searching rainfall metadata and generating rainfall product using the big data of nexrad. *Environmental modelling & software*, 117: 69–75, 2019.
- Seo, B.-C., Krajewski, W. F., and Ryzhkov, A. Evaluation of the specific attenuation method for radar-based quantitative precipitation estimation: Improvements and practical challenges. *Journal of Hydrometeorology*, 21(6):1333–1347, 2020.
- Sit, M., Demiray, B. Z., Xiang, Z., Ewing, G. J., Sermet, Y., and Demir, I. A comprehensive review of deep learning applications in hydrology and water resources. *Water Science and Technology*, 82(12):2635–2670, 2020.