Archives of Computational Methods in Engineering
https://doi.org/10.1007/s11831-021-09616-4

REVIEW ARTICLE

Big Data Analytics in Weather Forecasting: A Systematic Review

Marzieh Fathi[1] · Mostafa Haghi Kashani[2] · Seyed Mahdi Jameii[2] · Ebrahim Mahdipour[1]

Received: 27 December 2020 / Accepted: 12 June 2021

Abstract

Weather forecasting, as an important and indispensable procedure in people's daily lives, evaluates the alteration happening in the current condition of the atmosphere. Big data analytics is the process of analyzing big data to extract the concealed patterns and applicable information that can yield better results. Nowadays, several parts of society are interested in big data, and the meteorological institute is not excluded. Therefore, big data analytics will give better results in weather forecasting and will help forecasters to forecast weather more accurately. In order to achieve this goal and to recommend favorable solutions, several big data techniques and technologies have been suggested to manage and analyze the huge volume of weather data from different resources. By employing big data analytics in weather forecasting, the challenges related to traditional data management techniques and technology can be solved. This paper tenders a systematic literature review method for big data analytic approaches in weather forecasting (published between 2014 and August 2020). A feasible taxonomy of the current reviewed papers is proposed as technique-based, technology-based, and hybrid approaches. Moreover, this paper presents a comparison of the aforementioned categories regarding accuracy, scalability, execution time, and other Quality of Service factors. The types of algorithms, measurement environments, modeling tools, and the advantages and disadvantages per paper are extracted. In addition, open issues and future trends are debated.

1 Introduction

Originally weather forecasting started in the nineteenth century [1, 2]. The analysis of atmospheric data, including temperature, radiation, air pressure, wind speed, wind direction, humidity, and rainfall, is defined as weather forecasting. In order to predict the weather, a high volume of data must be collected or generated. Furthermore, these data are disorganized. Thus, utilizing the weather data for predicting the weather is a complex task, and it contains too many changeable parameters. These parameters vary according to the weather conditions that change very fast. To propose an algorithm for weather forecasting, we should consider its particular characteristics, such as continuity, data intensity, and multidimensional and chaotic behaviors [3, 4]. Originally weather forecasting has been developed from a human-intensive task [5] to a computational process [6], and to this end, it requires high-tech equipment. There are various factors that can affect the precision of forecasts. Season, geographical location, the accuracy of input data, classifications of weather, lead time, and validity time are some of these effective factors [7, 8].

Disorganized, heterogeneous, and enormous digital data are introduced as big data [9]. Utilizing conventional data management methods to process big data is not an easy and appropriate solution [10, 11]. To analyze this kind of data efficiently, we should search for a high-performance platform and a convenient big data mining algorithm to achieve advantageous information [9]. Big data quest operation for the purpose of disclosing hidden patterns, unknown relations, and other appropriate information, to make better decisions, which is called big data analytics [12]. In weather forecasting issue, big data is capable of upgrading the process of decision-making [13]. The importance of precise prediction can not be neglected; thus, we can use big data

✉ Seyed Mahdi Jameii
jamei@qodsiau.ac.ir

Marzieh Fathi
marzieh.fathi@srbiau.ac.ir

Mostafa Haghi Kashani
mh.kashani@qodsiau.ac.ir

Ebrahim Mahdipour
mahdipour@srbiau.ac.ir

[1] Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

[2] Department of Computer Engineering, Shahr-e-Qods Branch, Islamic Azad University, Tehran, Iran

Published online: 28 June 2021

analytics in weather forecasting to achieve this goal. Forecasting primarily was dependent on human forecasters, but now, in the age of information, it is done by applying technology and data [14]. Datasets of atmospheric data consist of rainfall, humidity, air pressure, radiation, sunlight intensity, data collection, etc. In addition, we need a huge number of datasets gathered from various sources (big data). Processing this amount of data requires high-tech hardware and software [15, 16].

Although weather forecasting highly depends on big data analytics, no systematic study is available to collect and evaluate the existing tools and techniques. Therefore, the key aim of our work is to review the approaches of big data analytics in weather forecasting. To this end, we have collected, arranged, and analyzed the existing approaches systematically. Our research mainly consists of the following tasks:

- Supplying a broad systematic survey of the existing approaches. In these approaches, big data techniques have been utilized to predict the weather.
- Proposing a technological classification according to current literature through the existing approaches.
- Studying the advantages, disadvantages, and evaluation types of the recent researches, rating evaluation determinants, and used tools in the proposed approaches.
- Considering the challenges of big data analytics in weather prediction
- Investigating the path for future research and the role of big data analytics in weather prediction

The remainder of the paper is structured as follows: the background of weather forecasting and big data analytics is presented in Sect. 2. Section 3 investigates the related surveys. Section 4 clarifies the research methodology. Also, Sect. 5 reviews the chosen big data analytics approaches in three classes. Section 6 outlines the findings and contrasts of techniques and technologies. Section 7 provides the open issues, and eventually, this research is concluded in Sect. 8.

2 Background

In this section, weather forecasting is discussed, and the central topic of big data is clarified. Lastly, the factors that have an impact on weather forecast performance are described.

2.1 Weather Forecasting

The daily atmospheric fluctuation is called weather. The weather data, like air pressure, temperature, speed and direction of the wind, humidity, etc., collected in various meteorological stations from sea observation, ground observation, radars observation, and other forms of observation show the current status of weather. These data are passed to several applications and models to discover various patterns [1], which is called weather forecasting. Weather forecasting originated in the nineteenth century [2], but due to the application of Numerical Weather Prediction (NWP), between the 1970s and 1980s, enormous progress has been made in weather forecasting operations [1]. Essentially, the mentioned models are made up of numerical equations. The computer program is used to solve these equations to monitor meteorological attribute alterations. It can be completed day by day or for some days in advance [17].

Weather forecasting is a vital process that is important for many people in their daily lives [18] and may impact many fields like agriculture, irrigation, and marine trade, and can save many lives from unexpected accidents [8]. Weather forecasting has many applications; there are many areas of life affected by weather forecasting, such as industry, transportation, disaster management, and energy management, that will be discussed below [14, 19]:

- Industry

A collection of manufactures or corporations that produce particular goods for sale and provide desired services is called an industry. The classification of the industry is based on the major source of income of a company or a group within that industry.

- Agriculture/Food

As the population grows, undoubtedly the need for food production increases. In order to enhance the quality of agricultural productions, weather forecasting and also big data analytics in this field should be used. Weather forecasting can help farmers be aware of soil erosion, overwatering, and drought. Predicting rainfalls makes farmers decide precisely on their crops, and they can also estimate the food price. Weather prediction also helps supermarket chains to arrange their stock control efficiently [16, 20].

- Tourism

Tourism refers to the act of individuals traveling to and staying somewhere away from the usual environment for pleasure, business, conferences, education, visiting friends, or other purposes. Choosing the destination for tourism is by far under the effect of climate conditions. In other words, the destination may be directly affected by climate change. So, the convenience and safety of tourists and their attractive targets are closely related to weather forecasting. Also, the benefits of this industry can be estimated by weather prediction [21, 22].

- Construction

Temperature, wind, humidity, and wet weather can certainly affect constructions. Thus, weather forecasting is very effective in protecting the work force, activities, and resources against the dangers of climate conditions. To put it differently, the sooner the threats posed by weather can be identified, the sooner the planning around them occurs. So, weather forecasting helps saving money and keeping the safety of the work force [23, 24].

- Sport

Sport industries are under the influence of such weather conditions as rainfall and lightning, so they employ weather forecasters; lightning is considered as a threat to the life of a golfer, and wind can highly affect the results of a sailing race [14]. When it rains, courts should be covered, so weather forecasting tells the time to do this task.

- Transportation

All over the world, weather forecasting helps shipping lines and crews to be aware of the dangers of storm, and thus, they can precisely decide on the time and route of sailing. Also, there may be delays or cancellations in flight schedules due to dangerous climate conditions; by means of weather forecasting, these problems can be solved [25, 26].

- Disaster management

There are natural disasters all over the world that are undeniable, but utilizing warnings and predictions can reduce the number of fatalities and save lives and also decrease economic damages. However, forecasting accuracy and its lead time are mainly based on the type of disaster. Tornado, flood, and landslide are some of these natural hazards that can be predicted by applying weather prediction in which big data analytics have been applied [27, 28].

- Energy management

Utilizing renewable energy is being dominantly far and wide. So, the need for accurate weather forecasting to gain the maximum information from sources of this kind of energy is unavoidable. Two of these sources are wind and solar energy, but the efficient use of these sources undoubtedly relies on weather conditions. Future intelligent power systems are possible by means of precise weather prediction [29–31].

There are two significant factors in weather forecasting: technological advances and human activities [32]. The improvement in communication capacity and information technology leads to the rapid enhancement of weather data that are both organized and disorganized data. There are systems that traditionally process organized data. These systems cannot process all disorganized data. This problem can be solved by utilizing big data analytics tools. As discussed earlier, precise and in-time weather forecasting can bring a lot of benefits to everyone in society. Improving weather forecasts is possible by using big data analytics [15].

2.2 Big Data

Relative to computational technology, the volume of data grows very quickly, and big data definitions are numerous. One of these concepts says that big data requires using new tools, analytics, and technical architectures to make high-value sources for businesses and to gain vast hidden information in analytics. Five main features characterize big data: variety, variability, volume, velocity, and value [33–36]:

- *Variety* comprises various types and formats of data. All the data from such different assets as sensor devices data, web log files, web pages, e-mail, documents, social networking sites are totally diverse.
- *Volume* is the extent of data and shows its immensities. Presently, the current data is in petabytes, and in the future, it is expected to be promoted to zettabytes.
- *Velocity* relates to the speed of the data changes, its creation rate, or the pace of the data transferring from several places. This feature relates to the data entry pace and also the pace of the data stream.
- *Variability* contemplates the incompatibility of the data stream. Maintaining the data loaded to the users' devices is very challenging, especially with the extension of social media.
- Data is of no use or validity in itself, but it requires to be transformed into priceless information, which is called *value*.

Three categories of big data, including structured, unstructured, and semi-structured data, are as follows:

- Unstructured data are not adapted to predefined data model, structure, and format. Therefore, it is comprised of data that are not easily searchable. Unstructured data include log data, X-ray images, CCTV footage, and videos [37].
- Semi-structured data have markers or tags to identify the logical components, but they do not match the normal patterns of data. They are characterized by partial, irregular, and implicit structure. The assemblage of various forms of structured data could lead to semi-structured data collection [37, 38].

- Structured data comprise layout, standard format, and extremely organized schema. Storing, accessing, and processing these data is in precise and organized ways. Modeling these data types needs a set of rules. Structured query language (SQL) is the highest frequency used tools in searching among structured data [37].

2.2.1 MapReduce

Google has generated MapReduce to process a big volume of data, such as online records and site application request reports on large service node clusters [39]. The MapReduce paradigm is an effective programming model for large-scale data-intensive computing applications [40]. Master, Map function and Reduce function are three parts of MapReduce. The Map and Reduce functions are managed by the Master and receive data and procedures from Master [39, 41]. Each job in a MapReduce application workflow creates Map and Reduce. Each input record is applied, and some intermediate records are produced by the Map function. Each bunch of these intermediate records with the same key is used, and some output records are produced by Reducer (the Reduce function) [42]. MapReduce model can run on a community of nodes on Hadoop [43, 44]. Therefore, all appropriate resources are managed by a master node to coordinate the contact between Mappers and Reducers. All separated parts of an input file, called input splits, are transferred to the Mappers, where they operate in parallel to supply the data found inside each split. As the Mappers supply the input, they divide the output; then, each Reducer collects the input partition from each Mapper, fuses and extracts them, and delivers the output file [42].

2.2.2 Big Data Tools

Big data can no longer be handled simply by using conventional methods and infrastructures, and research. Therefore, new techniques and methods designed for big data analytics and systems to store and handle these data are needed. Five important big data tools are introduced as follows:

- Hadoop

Hadoop system is a popular implementation of the MapReduce pattern. This tool enables applications to operate on extensive clusters that are constructed from generic hardware. Virtual Hadoop clusters over a huge physical cluster have been prepared by Apache Hadoop on Demand (HOD). The Hadoop MapReduce framework is composed of several slave nodes and a single master node. The master node executes an instance of Jobtracker that receives job requests from a client node. On the other hand, slave nodes execute a Tasktracker instance. The Jobtracker is a diagnostic tool contributing the status information to the client. Since it runs tasks in Java processes, several instances of the task can be completed in parallel [39, 45].

Hadoop Distributed File System (HDFS) has been created for storing huge files with streaming data access patterns and operating on generic hardware clusters. HDFS block size is 64 MB by default, much greater than the standard file system scale, and reduces the number of disk seeks. An HDFS cluster has two kinds of nodes: a number of datanodes called workers and a namenode called the master. The namenode has three tasks: the first one manages the file system namespace, the second one mainaines the file system tree, and the third one retains the metadata for all the files and folders in the tree. Storing and retrieving blocks are done based on the instructions of clients, or the namenodes are done by the datanodes. Datanodes report back the retrieved data with lists of storing blocks in the namenodes. However, HDFS will not work well in such cases: low-latency data access, lots of tiny files, multiple writers, and random file changes [33, 46].

- Apache Spark

Spark is an adaptable Hadoop data system and is a fast and general computing tool. Spark is executed in Hadoop clusters and processed all types of Hadoop data. Performing batch processing (like MapReduce) and new workloads such as machine learning, streaming, and interactive queries were the goals of designing Spark. By means of this engine, datasets are reinforced by in-memory processing. Hence the processing time will be improved. Scala is the language of Spark. The integration of Spark with Scala allows direct access and manipulation of datasets exactly just the same as local available objects. The difference between Hadoop and Spark is the possibility of parallel actions in Spark by utilizing datasets again and again. Also, access latency is decreased by cashing datasets in memory. Spark is optimized for machine learning algorithms by these features [39, 47].

- Apache Storm

The core of the system in big data real-time computing is Storm. Based on Eclipse Public License 1.0, Storm is a dispersed and fault-tolerance real-time processing system. Just the same as Hadoop manner in batch processing, Storm compiles and spreads difficult real-time computations in a computer cluster. Fast disposing of messages is assured by the use of Storm. Another advantage of storm is that its development is possible by using any programming language. Nimbus is a wonderful process executed on the main node for the aim of detecting errors, organizing tasks, and allocating codes. The supervisor is another wonderful process for working nodes to screen, begin, and halt the working process. Zookeeper is the subproject of Hadoop that handles the coordination of the work between Nimbus and Supervisor, and it accomplishes arranging the work in a large-scale dissemination system. Storm cluster and Hadoop are comparable as Nimbus and Supervisors can be compared to job tracker and task trackers, correspondingly [48].

- Apache Mahout

The aim of Apache Mahout is to contribute extensible and commercial machine learning methods to big data analysis applications. The core algorithms of Mahout consist of dimension reduction, pattern mining, regression, clustering, classification, and the like, which work in Hadoop platform control through the MapReduce model. These algorithms have good performances and capabilities because they are well-designed and optimized in the libraries. Mahout is designed to make a dynamic, receptive, different community to help discussions on potential use cases and on the project itself. In order to use Mahout, users should buy an Apache software license [49].

- Apache Kafka

Kafka is a distributed streaming platform. It handles streaming and operating data via in-memory computational strategies for real-time decision-making. Supporting distributed computing, concurrent data loading in Hadoop, and high-throughput are the major attributes of Kafka. The activity (clients' activities) and operational (servers' execution) data have been widely used in recent years to procure functional websites. Knowing how to interpret operating data is essential for controlling real-time activities. Kafka combines offline and online computations to contribute to real-time computing and provide an ad hoc approach for these two forms of data [44, 49].

2.3 Weather Big Data Analytic Factors

Each weather big data analysis approach has some evaluation factors containing QoS and weather factors for comparing the existing methods and solutions among recommended solutions. The important QoS and weather factors have been defined as follows:

- Accuracy, precision, and recall: In order to evaluate a big data method, to compare it with earlier methods to determine a more efficient algorithm, and also to discover its pros and cons, we need QoS factors. Precision, accuracy, and recall are some of these factors used in various articles. In the following, we represent the most important QoS factors for big data analytics in weather forecasting:

Accuracy, precision, and recall are defined as (1), (2), and (3). Accuracy is the percentage of correct prediction. The correct positive prediction percentage is called precision. The percentage of occurrences predicted as positive is called recall [50].

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \tag{1}$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{2}$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{3}$$

Additional factors used in articles for evaluation and comparison are execution time (second), the response time (second), scalability, reliability (a number between 0 and 1 that indicates the overall consistency of a measure), temperature (Celsius), wind (meters per second), precipitation (percent), humidity (percent), rainfall (millimeters), and pressure (Pascal).

3 Related Works and Motivation

Some researches about big data analytics have been done. This section presents some survey papers in the big data analytics field and the descriptions of their important factors and limitations.

Sahasrabuddhe and Jamsandekar [51] first introduced weather forecasting, including basic processes and different approaches, and then big data. They then explained a number of data structures used for big data and weather forecasting with an overview of a number of papers. This study did not have a classification, a research methodology, and a discussion, and this paper was not a systematic review.

Priya [52] described weather forecasting, big data, and rain forecast. The author reviewed several papers with the issue of predicting rainfall using data mining techniques. The author summarized them in a table and then began to expound the papers. The limitation was explained. This paper did not have a research methodology, the reviewed works, taxonomy, and discussion sections, so it was not a systematic review.

Hassani and Silva [53] carried out one of the big data analytics studies on forecasting. They introduced various kinds of challenges related to forecasting with big data, including noisy data detection, hardware and software, statistical significance, the architecture of existing algorithms, and big data itself. Also, the extant applications of statistical and data mining techniques for prediction were discussed. To provide the researchers with a more satisfactory experience, the authors summarized and categorized the reviewed studies based on the related fields, including economics, finance, population dynamics, crime, energy environment, biomedical science, and media. This paper was not a systematic review because it did not have a research methodology and a discussion.

Jain and Jain [14] discussed how such areas as the agriculture/food industry, tourism industry, sport industry, construction industry, transportation, disaster management, and energy were affected by weather forecasting. Then, they described technical challenges for weather forecasting, including managing large data sets, the availability of historical data, technological hurdles, the availability of forecast models, complexity, complex maintenance, and cost overrun. They did not describe the existing techniques and technologies of big data. This paper focused on weather forecasting areas rather than big data analytics; it was not systematic because it did not have a research methodology, the reviewed works, taxonomy, and discussion sections.

Reddy and Babu [15] planned to identify the problems in weather forecasting, survey various techniques, and to measure their efficiency. In this review, weather forecast methods were divided into very short-scale, short-scale, medium-scale, and long-scale forecast. They compared such different methods and models as the MapReduce model, linear regression method, random forest (RF), support vector machine (SVM), Neural Network (NN), Naive Bayes, REP tree and bagging algorithm, the nearest neighbors modeling, and wavelet artificial neural network (Wavelet-ANN) by analyzing 16 papers and investigating their data sets and hypothesis, the period of investigation, input parameters, methodology, results, and recommendations. This survey had a weak taxonomy and did not contain research methodology and result analysis. This paper was not a systematic review.

Rao [54] introduced climate changes and big data in the field of agriculture. The author then discussed big data technologies in agriculture, big data for climate smart agriculture, and a roadmap for leveraging big data for climate smart agriculture in India. This paper was not a systematic literature review because it did not have a research methodology, taxonomy, and discussion sections.

de Freitas Viscondi and Alves-Souza [55] systematically presented an article survey on the application of big data for solar photovoltaic electricity generation prediction. This paper included the most appropriate machine learning algorithms and significant papers in the solar forecasting field using machine learning. The authors described ANN, SVM, gradient boosting (GB), extreme learning machine (ELM), and random forest (RF); then, they analyzed ten key papers from 38 selected papers. The main drawback of this review was that the authors first answered some questions, performed a discussion of 38 selected papers, and then overviewed ten selected papers. They did not consider any QoS factors and did not apply any taxonomies in the mentioned approaches.

Mittal and Sangwan [17] introduced big data characteristics, technologies, and analytics. Then, they listed challenges in processing the extensive datasets by applying classic data mining algorithms and provided a literature review of 11 papers. They compared these papers in different weather prediction techniques such as KNN classification algorithm, MR-KNN and k-means algorithm on MapReduce platform, MapReduce and Linear Regression, and Bayesian model; they demonstrated the results in data set, techniques, parameters, and experimental results. Finally, they wrote the conclusion and future work. But no research methodology and taxonomy were presented in the reviewed approaches.

Vannitsem et al. [56] reviewed the approaches about statistical postprocessing of ensemble forecasts, the effect of utilizing statistical postprocessing, the problem of the impact of frequent model changes, and the use of blending techniques for correcting the forecasts. Then they evaluated the potential implementation difficulties of the statistical correction techniques. Finally, future prospects and challenges are discussed. This survey had a weak taxonomy and did not have a research methodology and result analysis. This paper was not a systematic review.

Considering the above-mentioned efforts, we recognized some defects in the previous works as follows:

- Most papers such as [14, 15, 17, 51–54, 56] are survey papers not SLR papers.
- The paper selection process was not clear in some papers such as [14, 15, 17, 51–54, 56].
- The QoS factors were not considered by researchers, and some of them, such as [15, 17, 52, 55], just considered the weather factors.
- Most papers such as [14, 17, 51, 52, 54, 55] did not classify the papers.
- None of the papers demonstrated clear statistical information on the modeling tools.

Table 1 demonstrates a list of the reviewed researches in which such parameters as review types, main topics, paper selection processes, publication years, future works, taxonomies, and covered years of each study are represented. It is obvious that only one systematic literature review paper and eight survey papers have investigated big data analytics for weather forecasting. Therefore, due to the declared reasons, we are excited to suggest a systematic literature review paper that improves all mentioned drawbacks, and our paper is the first review of the big data analytics in the weather forecasting field using the SLR method.

Table 1 Related studies in big data analytics for weather forecasting

References	Publication year	Main topic	Review type	Paper selection process	Taxonomy	Modeling tools	QoS and weather factors	Covered years
Sahasrabuddhe and Jamsandekar [51]	2015	Data structures used in big date and weather forecasting	Survey	Not clear	No	No	No	2012–2015
Priya [52]	2015	Using big data analytics to predict rainfall	Survey	Not clear	No	No	Weather factors	Not mentioned
Hassani and Silva [53]	2015	Using big data for forecasting	Survey	Not clear	Yes	No	No	Not mentioned
Jain and Jain [14]	2017	Weather forecasting applications and technical challenges	Survey	Not clear	No	No	No	Not mentioned
Reddy and Babu [15]	2017	Weather prediction models for early rainfall forecasting	Survey	Not clear	Yes	No	Weather factors	2011–2016
Rao [54]	2017	Utilizing big data for climate smart agriculture in India	Survey	Not clear	No	No	No	Not mentioned
de Freitas Viscondi and Alves-Souza [55]	2019	Big data models for solar photovoltaic electricity production prediction	SLR	Clear	No	No	Weather factors	2013–2017
Mittal and Sangwan [17]	2019	Weather prediction using big data analytics	Survey	Not clear	No	No	Weather factors	Not mentioned
Vannitsem et al. [56]	2020	Statistical post-processing techniques for weather forecasting	Survey	Not clear	Yes	No	No	2003–2020
Our paper	2021	Big data analytics in weather forecasting	SLR	Clear	Yes	Yes	Yes	2014-August,2020

4 Research Methodology

This section demonstrates a road map for a systematic review of the research relevant to big data analytic mechanisms in weather forecasting. There is a difference between a systematic review and a common traditional one: A systematic literature review (SLR) increases transparency and uses precise and replicable steps [57–62]. SLRs depend on clear and appraised review protocols to discover, analyze, and document results [63–65].

We designed a review process consisting of question formalization, paper selection, and documentation. Question formalization and paper selection steps are done in this section. For the documentation step, we categorize the approaches in Sect. 5 and then summarize the results in Sect. 6.

4.1 Question Formalization

The current work purpose is to answer the below-considered research questions (RQs) that are followed in the studies:

- RQ1: What is the importance of big data analytics in weather forecasting?
- RQ2: Which classification of the existing approaches can be used, and what are the advantages and disadvantages of this classification?
- RQ3: What are the applied QoS factors in big data analytics in weather forecasting, and what are the used weather factors?
- RQ4: What are the latest algorithms that help big data analytics in weather forecasting?
- RQ5: Which evaluation types are applied for evaluating the big data analytics in weather forecasting?
- RQ6: What are the utilized tools and frameworks for big data analytic approaches in weather forecasting, and how these are different?
- RQ7: What are the open issues and future trends of big data analytics in weather forecasting?

RQ1 was responded to in Sect. 2; RQ2, RQ3, RQ4, RQ5, and RQ6 are responded to in Sect. 6; and in Sect. 7, RQ7 is referred to.

4.2 Paper Selection Process

Figure 1 illustrates the selection process of the papers. According to Fig. 1, the paper selection strategy comprises three main steps:

Table 2 Online databases applied in paper selection

Online database	URL address
Google Scholar	http://Scholar.google.com
IEEE	http://ieeexplore.ieee.org/
ACM	http://dl.acm.org/
Science Direct	http://www.sciencedirect.com/
Springer	http://link.springer.com/
Taylor & Francis	http://www.tandfonline.com/
Wiley	http://onlinelibrary.wiley.com/
Inderscience	http://www.inderscienceonline.com
Sage	http://journals.sagepub.com.com
Emerald	http://www.emeraldgrouppublishing.com
World Scientific	https://www.worldscientific.com
Hindawi	https://www.hindawi.com

Step 1: Automatic search based on keywords, abstract, and titles.
Step 2: Selecting papers based on inclusion criteria.
Step 3: Selecting among the remained papers by reviewing the full text and elimination of the inappropriate papers.

In Step 1, we performed an electronic search on some famous academic databases such as Google Scholar, IEEE explorer, ACM, ScienceDirect, Springer, Taylor & Francis, Wiley, Inderscience, Sage, Emerald, World Scientific, and Hindawi, with the following related keywords:

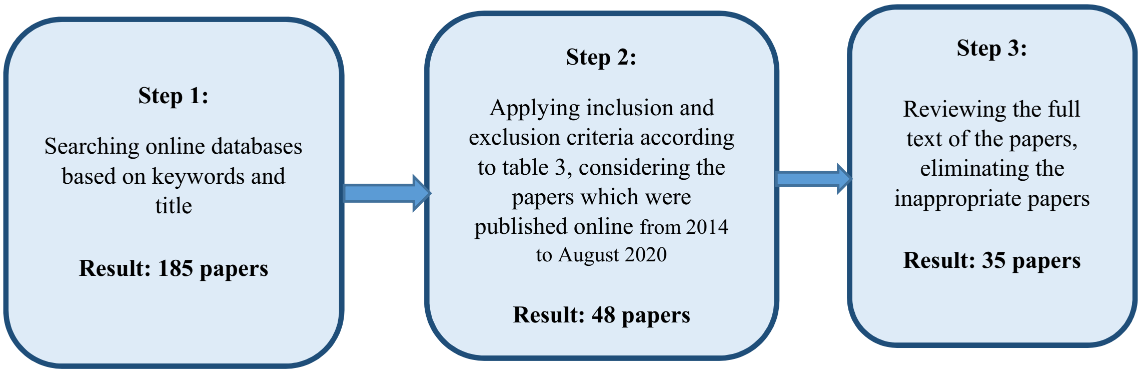
Step 1:
Searching online databases based on keywords and title
Result: 185 papers
Step 2:
Applying inclusion and exclusion criteria according to table 3, considering the papers which were published online from 2014 to August 2020
Result: 48 papers
Step 3:
Reviewing the full text of the papers, eliminating the inappropriate papers
Result: 35 papers

Fig. 1 Paper selection process

("big data" <OR> "large data" <OR> Hadoop <OR> Spark <OR> Storm)
[AND]
(weather <OR> meteorology)
[AND]
(forecasting <OR> prediction)

The search was performed in August 2020. As a result, we found 185 papers from journals, books, conference papers, notes, chapters, and any papers in English containing the mentioned keywords. The URL address of Electronic databases is shown in Table 2.

In Step 2, some inclusion and exclusion criteria according to Table 3 were considered. This step was taken for including only high-quality papers in the research. To this end, invalid journals and conference papers, white papers, reports, commentaries, editorial notes, and book review papers were omitted. Then 48 papers remained.

In Step 3, the whole texts of the papers were reviewed to check if the papers were relevant to the subject. After eliminating the inappropriate papers, 35 papers were selected.

Table 3 Inclusion/exclusion criteria

Criteria	Justification
Inclusion	
The studies suggesting evaluations or solutions for big data analytic mechanisms in weather forecasting	The purpose of these studies was to probe some solutions for big data analytic mechanisms in weather forecasting
ESCI-indexed journals and conference papers	These papers provided a high level of quality
Studies published online from 2014 to August 2020	The term big data had emerged from 2005 when O'Reilly Media launched it. The researches about weather big data have been started since 2012
Exclusion	
Studies on big data not focusing on weather forecasting	The aim of this research was to investigate solutions for big data analytic mechanisms in weather forecasting, so any study not including this issue would be excluded
Thesis, books, or book chapters	The researches in this category were related to journals or conference papers most of which were considered in our paper
Review papers or commentaries	These types of papers suggested no creation or explicit solution
Non-peer-reviewed papers	Due to the uncertainty about the quality of the unjudged papers, these papers were excluded

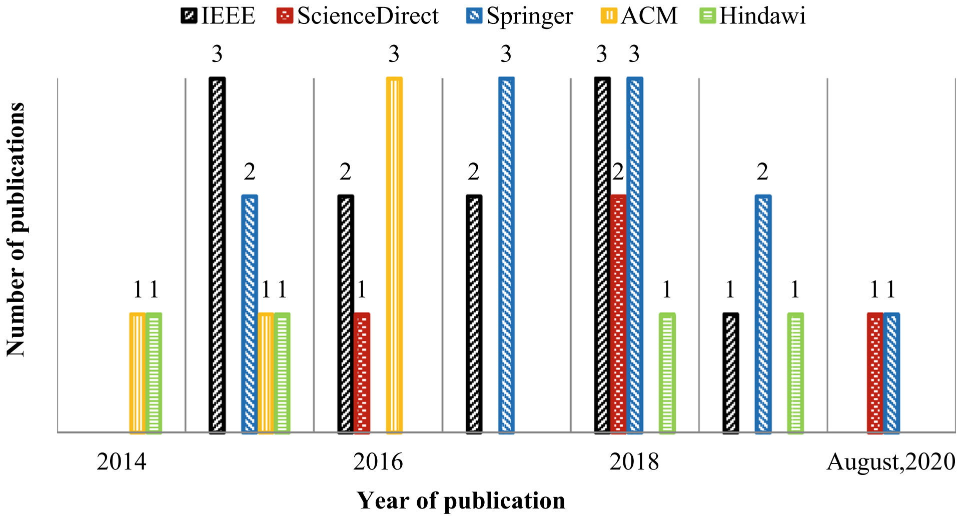

Fig. 2 Distribution of papers per publisher

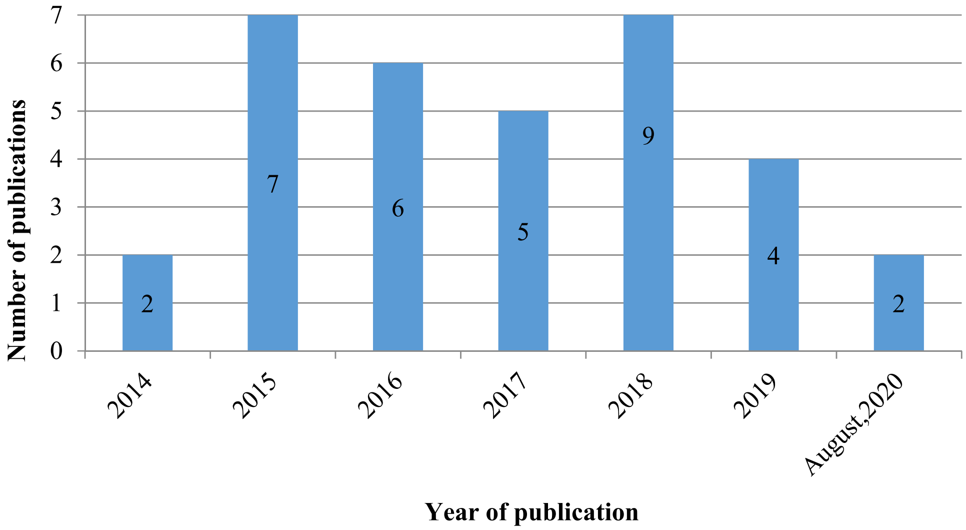

Fig. 3 Number of publications per year based on Step 3

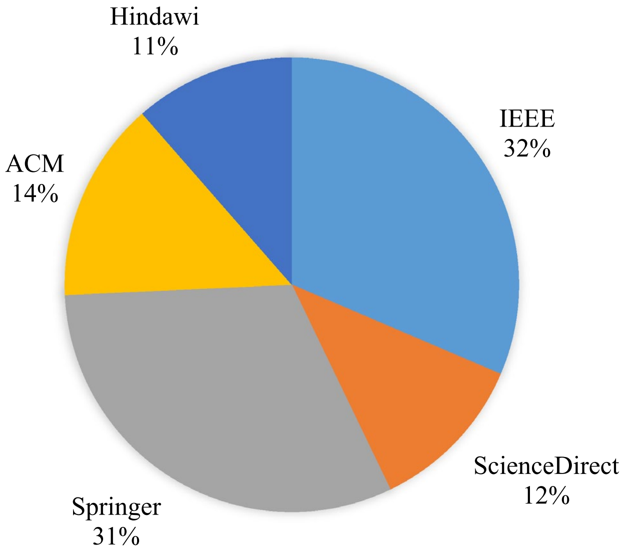

Fig. 4 Percentage of the big data analytic mechanisms in weather forecasting papers among different publishers based on Step 3

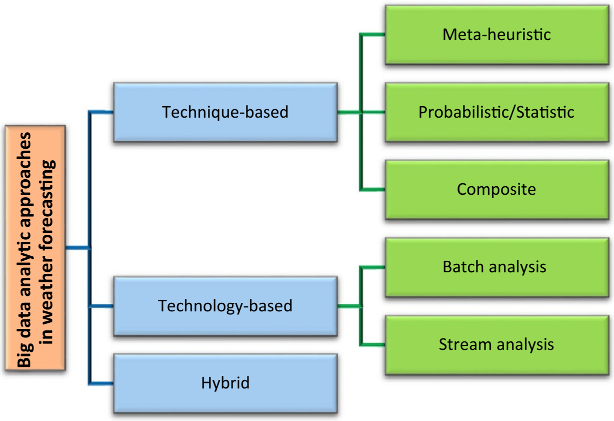

Fig. 5 Taxonomy of big data analytic approaches in weather forecasting

The distribution of these 35 papers per publisher is depicted in Fig. 2.

The number of big data analytics and weather forecasting publications in each year, from 2014 to August 2020, is pointed in Fig. 3. This figure pointed that most of the papers in our reviewed field were published in 2018.

Figure 4 illustrates the classification of the papers among five publishers from 2014 to August 2020, where 32% of the papers were related to the IEEE, the same percent of papers belonged to Springer (31%), 14% of the papers belonged to the ACM, ScienceDirect published 12% of these papers, and 11% of the papers were related to Hindawi.

5 Classification of the Selected Approaches

In this section, 35 chosen papers are surveyed considering the indicated criteria. So, defining and discussing the techniques and the basic properties of each paper along with their variations, advantages and disadvantages are presented. As the literature presented on big data and weather forecasting is so diverse, the organization or review of the studies is never an easy task. Based on the studies selected, it was understood that all of them could be related to technology and techniques or a mixture of the two. Therefore, the surveyed papers can fall under common umbrellas. The researchers focused on big data analytics by regarding three main categories, including technique-based, technology-based, and hybrid approaches. Technique-based approaches are categorized as meta-heuristic, probabilistic/statistic, and composite methods. Technology-based approaches are based on big data ecosystems, tools, frameworks, etc., and mostly try to scale up and speed up weather data analysis to improve weather forecasting. These approaches are classified as batch analysis and stream analysis. In the batch processing methods, a high volume of repetitive data can be processed all at once with little or no user interaction. In these methods, at first, data are collected over time and then are sent for processing. In the streaming methods, data are sent for processing piece-by-piece, and this processing is usually done in a real-time manner. A hybrid approach is a combination of big data techniques and technologies. Figure 5 illustrates the classification of approaches. In Sects. 5.1, 5.2, and 5.3, these approaches and their examples are discussed. In addition, reviewing each classification is done based on the used evaluation types, tools, advantages, disadvantages, and algorithms.

5.1 Technique-Based Approaches

A technique-based approach in big weather data analytics often focuses on the quality of the forecasts, which depends on the data quality in several levels of its life cycle. Most of these approaches try to discover the hidden knowledge in the observed data. In Sect. 5.1.1, the discussion of the chosen technique-based approaches are categorized as meta-heuristic [66, 67], probabilistic/statistic [68–79] and composite [50, 80] methods is performed, and in Sect. 5.1.2, their outlines are displayed. The technique-based approach is the most utilized approach in weather forecasting methods.

5.1.1 Overview of the Chosen the Technique-Based Approaches

Cheng, et al. [66] proposed attribute reduction based on a genetic algorithm for the co-evolution of meteorological data (AECMD). The crossover operator and mutation operator of the adaptive prediction algorithm were developed, and the evolutionary population was split into two subpopulations: the first one improved the convergence speed by using the elite-assisted cross-operation, and the second subpopulation maintained the population diversity within the evolutionary process by introducing a random population, and the two subpopulations completed the iterative operation. AECMD algorithm was compared with the Tabu Discrete Particle Swarm Optimization (TSDPSO) and Attribute Reduction algorithm based on the Adaptive Genetic Algorithm (ARAGA) algorithms. Also, Cramer et al. [67] recommended a hybrid Genetic Programming/Genetic Algorithm (GP/GA) algorithm to increase the predictive accuracy of rainfall. The overall problem was decomposed into a collection of partitions, where the GP component of the GP/GA created multiple regression equations. By implementing the GP, it was expanded with the proposed GA classification variable to decide which regression equation to appraise. The authors compared the performance of the GP/GA against Markov-Chain extended with Rainfall Prediction (MCRP), Radial Basis Function (RBF), and GP without decomposition.

In order to discover the likeness among the weather factors with the highest accuracy of feature selection, Pooja et al. [68] applied Tanimoto correlation coefficient. Next, the MAP expected clustering procedure was performed to divide the weather data into groups for cluster formations. To improve the performance of prediction, the clustering outcome was given to the linear program boosting classifier. The results evidenced that, compared with routine methods, the TC-CMECLPBC technique increased the Prediction Accuracy (PA) with lower False Positive Rate (FPR) and time. Further, Kvinge et al. [69] proposed the kappa-profile, statistical analysis, and monitoring of large datasets, arising from a problem of dimensionality-reduction optimization. Computationally, possible means of extracting kappa from huge datasets were provided by the Secant-Avoidance Projection (SAP) and the Hierarchical Secant-Avoidance Projection (HSAP) algorithms. The comparison of the returned projection of the SAP algorithm with the popular dimensionality-reduction method of Principal Component Analysis (PCA) was once executed by the concept of geodesic distance on a Grossmann manifold. In order to obtain an even stronger performance of forecasts, Buszta and Mazurkiewicz [70] suggested the concept of utilizing visualization to prefilter the data and move it on to a neural network. Visual investigation permitted making decisions on areas that were related to data locations that should be accustomed to training, validating, and testing neural networks. The selected neural network model was a multilayer perceptron.

In addition, Rasel et al. [71] used SVR and ANN to compare the performance of data mining and machine learning methods for strong weather forecasting goals. SVR could outdo the ANN in rainfall prediction with a negligible error rate using all dataset forms, and ANN could achieve stronger outcomes than the SVR with a reasonable deviation of error rate. The Cumulative Distribution Function (CDF) was proposed by Mahmood et al. [72] for modeling and analysis of complicated data for weather forecasting. They studied the Chattisgarh area in India because of the unpredictability of the specific climate shift in its atmosphere. The authors inspected two events on various seasons with specific variables and areas. The weather forecasting for the next three years could be effectively done by this strategy. Also, Azimi et al. [73] proposed T.S.B K-means to make it more suitable for neural networks by classifying data into different groups and identifying anomalies and abnormal patterns. The Discrete Wavelet Transform (DWT) disintegrated the wind power data; the Harmonic Analysis of Time Series

(HANTS) filtered DWT, and finally, it was pre-processed to provide the inputs for the Multilayer Perceptron Neural Network (MLPNN). The cluster selection method decided the maximum correlation data cluster with testing data to furnish the training data for the wind strength forecasting in every individual hour.

For multi-label classification of weather, a deep learning model was tested by Doreswamy and Manjunatha [74], and Long Short-Term Memory (LSTM) and fully related models were combined. In this work, the deep LSTM model overcame the data mining method in the field of weather classification. While the suggested model was implemented well on majority class labels, it should also be comparable to minority class labels. Also, Venkatachalapathy et al. [75] proposed a framework for weather data classification and prediction by the combination of C5.0 and Support Vector Clustering (SVC) algorithms. The performance of the two classifiers was evaluated by using failure ratio, success ratio, gain, and standard deviation. The future weather forecast was calculated and saved as a virtualization of the dataset and then transferred simultaneously to n number of clients using the network interface unit. Also, Choi et al. [76] created a model for the heavy rainfall prediction. Two algorithms were derived that predicted damage from heavy rainfall on a determined day. They used data of the same-day weather observation in the first algorithm and data of the past weather observation in the second algorithm. Each of these algorithms was implemented in machine learning (boosting, random forest, bagging, decision tree). The second algorithm was determined to be an alternative to the first algorithm, and the final model was a boosting model applying past weather data because of achieving the best performance of prediction between second algorithm models.

Furthermore, Hubig et al. [77] presented a framework to predict natural hazards. In order to discover local outliers in the space–time- and attribute-space, an Environmental Extremeness Measure (EEM) score was recommended. In addition to the unsupervised learning approach, a join of sparsifying data on the basis of EEM score, developing an environmental sensor, and the application of Coupled Tensor-Tensor Factorization (CTTF) or Four-Mode Tensor Factorization (4MTF) on it, and categorizing or forecasting this data was developed as a model. A data mining methodology was submitted by Yonekura et al. [78] on the basis of dense weather station device for short-term weather forecasting. The idea consisted of two folds: the model for point prediction and the model for tensor prediction. This model outperformed XGBoost and supported vector machines using big real observation data. Based on experiments, deep neural networks yielded the highest accuracy for rain prediction, and in the case of "No-Rain" prediction, the block-type network had better accuracy than the fully connected one. The bad data of Numerical Weather Prediction (NWP) was mined, and a short-term Wind Power Forecasting (WPF) model was proposed by Xu et al. [79]. They introduced a bad data analyzer to thoroughly find out the connection between the WPF error with the raw NWP extracted features. Then, a hierarchical structure, composed of a K-means clustering-based bad data detection module and a Neural Network (NN)-based forecasting module, was suggested. Finally, the irregular raw NWP data was regulated before despatching the WPF model.

Lee et al. [50] applied Genetic Algorithm (GA) for finding a far smaller weather data collection without performance reduction. The first step was evaluating the weather parameters by using SVM. The next step was applying GA for choosing the correct features for improving the computational cost and accuracy of prediction. As a result, the accuracy of prediction, applying the original data and dimension-reduced data, was alike, but SVM computation time decreased nearly eight times. In another work, Jiang and Dong [80] proposed an intelligent hybrid model to tackle wind speed forecasting. Initially, they used Wavelet Transform (WT) for taking out the key components of the initial wind speed data while disposing the clamor. The combination of the back-propagation artificial neural network and the Artificial Fish Swarm Algorithm (AFSA) was applied to predict wind speed.

5.1.2 Summary of the Technique-Based Approaches

Table 4 displays more details of the studied papers, including technique type, main idea, evaluation type, tool, and advantage and disadvantage used in the selected techniques, as well as their advantages and disadvantages. Most of the studied papers used Non-heuristic algorithm and machine learning as their evaluation types and different tools for modeling the approaches. The main advantage of the technique-based approach is gaining high accuracy, while high execution time is its disadvantage. The existing evaluation factors in technique-based approaches are shown in Table 5.

5.2 Technology-Based Approaches

A technology-based approach, which is based on big data ecosystems, tools, frameworks, etc., mostly tries to scale up and speed up weather data analysis in order to improve weather forecasting. In Sect. 5.2.1, the chosen technology-based approaches are classified as batch analysis [81–86] and stream analysis [87–89] are described, and their summaries are conferred in Sect. 5.2.2.

Table 4 Categorization of recent papers in the technique-based approaches

Technique	References	Main idea	Evaluation type	Tool(s)	Advantage(s)	Disadvantage(s)
Meta-heuristic	Cheng et al. [66]	Meteorological data attribute reduction based on a genetic algorithm	Data sets	Not mentioned	Low execution time High reduction performance Avoiding premature convergence	Problem with coding the fitness (evaluation) function
	Cramer et al. [67]	Proposing a hybrid genetic programming/genetic algorithm (GP/GA) for decomposing rainfall	Data sets	IRace package	High classification performance than GP and MCRP (low RMSE)	No testing other regression and classification techniques Unable to create complex rule No specifying the partitions through time
Probabilistic/statistic	Pooja et al. [68]	Enhancing the weather prediction accuracy with minimal time by TC-CME-CLPBC Technique	Data sets	Not-mentioned	High prediction accuracy Low execution time	
	Kvinge et al. [69]	Dimension-driven data analysis using a statistic dimensionality-reduction technique called the kappa profile	Formal	Not mentioned	High effectiveness (capturing the dimension of non-linear data)	Low reliability Unable to estimate the minimum embedding dimension in the underlying dynamic system
	Buszta and Mazurkiewicz [70]	Weather forecasting using visual data presentation techniques and neural networks	Data sets	Isopleths	Easy access to data Easily parametrized High possibility to forecast multiple meteorological events High scalability (multiplicity of parameters)	Low ability to detect temperature extremes Manual location choice Human interaction needs
	Rasel et al. [71]	Comparing the performance of SVR and ANN for a robust weather prediction purpose	Data sets	Not-mentioned	High prediction accuracy	High execution time by using SVR
	Mahmood et al. [72]	Forecasting climate change based on data mining techniques	Formal	MATLAB	High prediction accuracy for the next three years	Low scalability
	Azimi et al. [73]	Accurate wind power forecasting based on data mining	Data sets	Not mentioned	Low execution time	Low prediction accuracy
	Doreswamy and Manjunatha [74]	Building a high accuracy multi-label classification model	Data sets	Theano	High classification accuracy	Unable to predict minority class labels Execution time was not evaluated

Table 4 (continued)

Technique	References	Main idea	Evaluation type	Tool(s)	Advantage(s)	Disadvantage(s)
	Venkatachalapathy et al. [75]	Presenting a weather data classification and prediction framework based on data mining techniques	Data sets	Apache Hadoop	Low expected misclassification cost High memory efficient than C.4.5 Low execution time High ability for boundary formation	Not supported real-time dataset
	Choi et al. [76]	Developing a heavy rain damage prediction model using machine learning	Data sets	Not mentioned	High prediction performance	Low learning scalability of boosting
	Hubig et al. [77]	Detecting and predicting natural hazards by machine learning approach	Data sets	MATLAB	High prediction accuracy	Low ability to outperform the proposed baseline classification for other classification algorithms
	Yonekura et al. [78]	Short-term weather forecasting based on the tensor prediction model	Data sets	Scikit-learn Python xgboost module	High prediction accuracy	Low prediction accuracy in the lengthened horizon
	Xu et al. [79]	Mining the bad data of numerical weather prediction (NWP)	Simulation	Not mentioned	Low total WFO error	Unclear reason of different error patterns Low reliability
Composite	Lee et al. [50]	Finding out the faster weather data feature selection method for heavy rain forecasting based on genetic algorithm	Data sets	SVM Light	Low execution time Low computational cost High effectiveness	Not improving the prediction accuracy Low simplicity Long time to find the appropriate solution using GA
	Jiang and Dong [80]	Wind speed forecasting based on artificial neural networks	Simulation	MATLAB	High prediction precision	High execution time Less efficient discretization of WT

Table 5 Comparison of the existing evaluation factors in the technique-based approaches

Technique	References	Temperature	Time	Wind	Accuracy	Scalability	Humidity	Precipitation	Rainfall data	Pressure	Mean square error	Reliability	Root mean square error	Precision
Meta-heuristic	Cheng et al. [66]	✓	✓	✓			✓	✓		✓				
	Cramer et al. [67]								✓				✓	
Probabilistic/Statistic	Pooja et al. [68]	✓	✓		✓									
	Kvinge et al. [69]											✓		
	Buszta and Mazurkiewicz [70]	✓		✓		✓				✓				
	Rasel et al. [71]	✓	✓		✓				✓				✓	
	Mahmood et al. [72]	✓			✓	✓								
	Azimi et al. [73]	✓	✓	✓	✓									
	Doreswamy and Manjunatha [74]				✓						✓		✓	✓
	Venkatachalapathy et al. [75]		✓											
	Choi et al. [76]	✓		✓		✓	✓	✓		✓				
	Hubig et al. [77]	✓		✓	✓			✓						
	Yonekura et al. [78]	✓		✓	✓		✓	✓	✓	✓	✓			
	Xu et al. [79]											✓	✓	
Composite	Lee et al. [50]	✓	✓	✓	✓		✓							✓
	Jiang and Dong [80]	✓		✓							✓			✓

5.2.1 Overview of the Chosen Technology-Based Approaches

More et al. [81] demonstrated that the analysis of weather data by using Hadoop with a Multiple node HDFS system was required due to the increasing volume of the weather data. MapReduce has the potential to discover hidden weather patterns. In order to benefit from Hadoop's parallel operation, Wu [82] represented a method to make an analysis of the weather log swiftly. Hadoop platform adapted the core technology of HDFS and MapReduce, joining with Java Web's programming techniques and concepts, formed a B/S structure, and used the plug-in high charts for producing charts. Also, Ismail et al. [83] proposed big data prediction framework on the basis of the MapReduce algorithm for weather temperature. For the smooth and effective analysis of a wide range of data, they used Apache open-source Hadoop, including the HDFS and Hadoop MapReduce, besides other components that prepared high-speed clustered processing. The experimental system was built on a cluster with three PCs. In addition, Abdullahi et al. [84] conducted the performance profiling of Meteorological and Oceanographic Data (MOData) on Hive for analyzing, loading, and formatting. The formatting was done via tailored Serialization and De-serialization (SerDe). The loading and fine-tuning were reached with bash script and partitioning to a grainy. In contrast to the usual database systems, the results presented that if MOData excellently configured its analytics with Hive, it would be clearly successful. Also, Oury and Singh [85] used Hadoop to analyze the weather data on climate change for predicting rainfall. For MapReduce operation, they utilized Apache PIG to perform the Extract, Transform, and Load (ETL) acts, so all missing and incomplete data were found out and deleted. Then, to access the library and actions like parsers, Python programming language was utilized. Running three Linux SED commands resulted in the manipulated data compatible with the JSON pattern. Furthermore, Manogaran et al. [86] proposed a framework to process big data for integrating health and climate data and finding the mutual relation among the climate factors and dengue occurrence. This framework consisted of the MapReduce programming model, Hive, HBase, and ArcGIS that were presented in a five layers HDFS environment, including the data ingestion layer, batch layer, speed layer, serving layer, and visualization layer.

Jayanthi and Sumathi [87] presented a spark-based technique to analyze weather data. A Spark instance with an Ipython notebook on it was created. Then, a number of partitions or Resilient Distributed Datasets (RDDs) were distributed in a cluster, and the data was created into them for executing Spark streaming. RDD transformation operations were used to analyze the collected meteorological data. Eventually, for the first ten meteorological stations, the maximal average precipitation and maximal average temperature were calculated and shown. Also, Palamuttam et al. [88] presented a big data framework based on Apache Spark (SciSpark). In this study, the parallel Ingestion and partitioning of satellite and model datasets for earth science were demonstrated. The authors implemented the facet of GTG (a method for identifying and tracking mesoscale convective complexes in satellite infrared datasets) algorithm by using SciSpark and its MapReduce. To fill in user-defined issues, the architecture of SciSpark upheld the benefits of the increasing distributed and sequential programming. Finally, Hassaan and Elghandour [89] proposed a system (DAMB) consisting of Kafka and Spark streaming for reading streamed data and a heterogeneous cluster of CPUs and GPUs for real-time processing of such data. SparkGPU was DAMB's kernel, the Apache Spark extension platform. They implemented two approaches of SparkGPU: Java Native Access (JNA) and node manager.

5.2.2 Summary of the Technology-Based Approaches

Table 6 displays more details of the studied papers, including technology type, main idea, evaluation type, tool, and advantage and disadvantage used in the selected techniques as well as their advantages and disadvantages. The real testbed is the mainly used evaluation type; big data tools are used for modeling the approaches, and most of them did not use any algorithms. The main advantage of the technology-based approaches is gaining low time, including execution time and response time, and their main disadvantage is low reliability. The technology-based solutions for big data analytics in weather forecasting are utilized to improve scalability and time consumption via applicable approaches. The existing evaluation factors in technology-based approaches are shown in Table 7.

5.3 Hybrid Approaches

In Sect. 5.3.1, we review the selected hybrid approaches. Commonly, complex predictive analysis cannot be done by a basic approach, so two of them are merged to gain a value-added approach. A hybrid approach is a combination of big data techniques and technologies. In Sect. 5.3.2, we summarize the main ideas of the reviewed papers.

5.3.1 Overview of the Chosen the Hybrid Approaches

A climate change detection algorithm based on spatial cumulative sum (CUSUM) was proposed by Manogaran and Lopez [90]. The CUSUMs of each sample value deviation from the target value (average) were determined by a CUSUM control chart. The large day-wise climate data was stored in a distributed manner on HDFS. In order to

Table 6 Categorization of recent papers in the technology-based approaches

Technology	References	Main idea	Evaluation type	Tools	Advantage	Disadvantage
Batch analysis	More et al. [81]	analyzing weather data impressively by using Hadoop with MapReduce	Real testbed	Apache Hadoop	Better accuracy than existing systems Low execution time	High processing Overhead due to using Hadoop
	Wu [82]	Analyzing the weather log and taking advantage of the parallel process on Hadoop	Real testbed	HBase Apache Hadoop	Low execution time	Low ability to handle small files
	Ismail et al. [83]	Representing analytical Big Data prediction framework for weather temperature based on MapReduce algorithm	Real testbed	Apache Hadoop	High scalability (Removing scalability bottleneck) High effectiveness (analyzing the temperature effectively) Low execution time (faster processing of the data)	Low simplicity (Not Easy to Use) Any caching Low security
	Abdullahi et al. [84]	performance profiling of Meteorological and Oceanographic data on Hive	Real testbed	Apache Hive	Low response time HiveQL is a declarative language like SQL High efficiency	Useful when the data is properly formatted Debugging code is very difficult
	Oury and Singh [85]	Analyzing the climatic weather data using Hadoop Technology	Real testbed	JQueryFlot Library Apache Pig Python	Huge parallel handling capacities Low development time	High execution time
	Manogaran et al. [86]	Processing the big climate data using big data technologies	Real testbed	Apache HBase Apache Hive ArcGIS 10.2	Low response time	Unable for real-time computing
Stream analysis	Jayanthi and Sumathi [87]	weather data analysis based on spark implementation	Real testbed	Spark Streaming MLib Kafka Flume Ipython notebook Matplotlib	High scalability High fault tolerance Low execution time Low disk I/O	Computational cost caused by Spark was not evaluated
	Palamuttam et al. [88]	Presenting a distributed In-memory framework for detecting weather event	Real testbed	Apache spark PySpark	Low execution time High scalability (ability to process high-resolution grids) Low disk I/O Maintaining the chronological order necessary for the graph creation	High network latency Tightly coupled reading hierarchical files with the HDFS Big data problem with Cartesian product
	Hassaan and Elghandour [89]	Processing streamed weather data on a heterogeneous cluster of CPUs and GPUs in real-time	Real testbed	Kafka Spark streaming SparkGPU	Low execution time Low communication time	Low scalability Low reliability

Table 7 Comparison of the existing evaluation factors in the technology-based approaches

Technology	References	Temperature	Time	Wind	Accuracy	Scalability	Humidity	Precipitation	Rainfall data	Reliability
Batch analysis	More et al. [81]	✓	✓		✓				✓	
	Wu [82]	✓	✓							
	Ismail et al. [83]	✓	✓			✓				
	Abdullahi et al. [84]		✓							
	Oury and Singh [85]	✓	✓					✓	✓	
	Manogaran et al. [86]	✓	✓	✓			✓	✓		
Stream analysis	Jayanthi and Sumathi [87]	✓	✓			✓		✓		
	Palamuttam et al. [88]	✓	✓			✓				
	Hassaan and Elghandour [89]		✓			✓				✓

calculate the seasonal average of climate parameters, the MapReduce framework was applied to that data. The authors compared the above-mentioned climate change detection algorithm with segment neighborhood, binary segmentation, Pruned Exact Linear Time (PELT), and CUSUM with Bootstrap methods. Also, in Madan et al. [91], progressive statistical linear regression and SVM, using Hadoop, were outflanked by weather detecting directions, despite diminishing their executing mistakes some days later approximately. In their proposed framework, the authors had an inclination to infuse the enhanced rule of the algorithm that provided approximate outcomes to predict the climate for five days later. Results were computed by the mathematical and statistical decision tree and confusion matrix concepts, utilizing big data to forecast more accurately.

Dhoot et al. [92] performed a comparative analysis according to the weather prediction calculation time and mean square error with and without the Spark cluster for the ARIMA model and Kalman Filter. In this paper, the single performances of the ARIMA Model and Kalman Filter were analyzed. For classifying the weather condition by means of values forecasted by the models, XGBoost Classifier was used. ARIMA model with Spark demonstrated important improvement in efficiency in terms of run time. In another work, Pandey et al. [8] processed the unstructured weather data using the Hadoop framework. Word count algorithm was used for pre-processing the stored data through Hadoop. After that, they achieved the final dataset for the weather prediction process. Two data mining tools were applied for the weather prediction: Artificial Neural Network Fuzzy Interface System (ANFIS) and Fuzzy Logic (FL).

Bendre et al. [16] proposed a solution for the problem of discovering supplementary insights from exact data on agriculture via a big data approach. The objective was to enhance the precision of prediction by using various weather factors for future precision agriculture. The authors explained the findings by applying a programming model and a distributed weather application algorithm for data processing and forecasting. The suggested predictive model contained the MapReduce for big data processing and linear regression for data forecasting. Also, Dhamodharavadhani and Rathipriya [93] proposed Region-wise rainfall forecasting by means of MapReduce-based exponential smoothing methods, including holt's linear, simple exponential smoothing, and holt-winter's exponential smoothing. As the results suggested, Holt-Winter's Exponential Smoothing showed better precision and performance. Using MapReduce demonstrated major enhancements in runtime in comparison with serial implementations. In conclusion, and based on the results, MapReduce-based Holt-Winter's smoothing method was the best method for region-wise monthly rainfall prediction.

Namitha et al. [94] implemented ANN on the MapReduce framework for short-term (next day's) rainfall prediction.

Three types of forecasting—intensity prediction (classification), rain/no-rain prediction (classification), and regression model using neural networks- were compared in two ways: non-parallelized and MapReduce approach. Even for the great data size, e.g., terabytes or petabytes, the same framework worked well for rain forecasting. But only the efficiency of regression enhanced in batch learning.

Liu et al. [95] proposed a Frequent Pattern (FP)-tree algorithm based on MapReduce to predict weather conditions in a cloud computing system. They used the map operation to divide the meteorological data mining issue into several blocks for making the intermediate key/value pairs set and the reduction operation to merge the above values to make as small a set of values as possible. It was observed from the experimental results that, by enhancing the number of computing nodes, the algorithm run time decreased apparently.

Sahoo [96], the proposed framework contained saving NCDC semi-structured data on the Hadoop cluster and applying a clustering methodology for weather forecasting. For designing clusters demonstrating the relation between the current and previous year's weather data, conventional k-means clustering was used. For processing the current year's weather parameters, they used an incremental k-means clustering algorithm and found that the computed weather condition occurred in one of the existing clusters to express similar atmospheric conditions. Fang, et al. [97] suggested a developed parallel *K*-means (MK-means) algorithm on the basis of MapReduce. They adapted the *K*-means algorithm in Hadoop, and applied *MK*-means for categorizing the extensive weather data. Then, a rapid *MK*-means clustering algorithm was proposed to analyze weather information processing by using MapReduce. It was found that the proposed *MK*-means algorithm applied in the big weather data processing system was efficient.

5.3.2 Summary of the Hybrid Approaches

The hybrid approaches for big weather data analysis were utilized, and both mentioned approaches were combined. So hybrid approaches scale up and speed up weather forecasting mechanisms and improve the quality of model outputs. A side-by-side comparison of the chosen hybrid approaches is illustrated in Table 8 shows more information about the studied papers, including the main idea, evaluation types, tool, and advantage and disadvantage used in the selected techniques as well as their advantages and disadvantages. The real testbed is the mainly used evaluation type, big data tools are used for modeling the approaches, and most of them used non-heuristic algorithms. The main advantage of the technology-based approaches is gaining low execution time. The main disadvantage of the *technology*-based approach is low reliability. Also, the existing evaluation factors in the hybrid approaches are shown in Table 9.

6 Discussion

Section 5 described the reviewed papers in the field of big data analytics in weather forecasting. In this section, the qualitative and quantitative analysis of the presented reviews is discussed. In addition, we answer the research questions mentioned in Sect. 4.1 as follows:

- **RQ2**: Which classification of the existing approaches can be used, and what are the advantages and disadvantages of this classification?

According to Fig. 6, the technique-based approach of the three approaches described in Sect. 5 had the greatest number of published papers with 46%. Other approaches, hybrid, and technology-based approaches, with 28% and 26%, came next. Also, Table 10 presented a summary of the pros and cons of the discussed classes extracted from Tables 4, 6 and 8. It demonstrated that technique-based approaches provided high prediction accuracy, low execution time, and low computational cost and mostly had low reliability, low scalability, and low simplicity. The technology-based approaches, in comparison with the above-mentioned approaches, were the same in execution time, and in addition, they had high scalability. However, these approaches caused high network latency and low simplicity. The hybrid approaches achieved all of the advantages of the two approaches mentioned above, but they had low reliability as a disadvantage.

- **RQ3**: What are the applied QoS factors in big data analytics in weather forecasting, and what are the used weather factors?

According to Figs. 7, 8, 9, and 10 that visualized the results mined from Tables 5, 7 and 9, most researchers have explained different factors, although others have ignored them. Based on the factors mentioned in Sect. 5 divided by taxonomies, it is prominent to find the most important evaluation factors, including QoS and weather factors. To grasp the importance percentage of each factor, we used (4), the number of happenings of a factor i (ocurr_no(i)), numerated individually, and divided them by the sum of the number of happenings of all factors. The importance percentage of factor i (Imp_percentage(i)) is obtained by multiplying its calculated value by 100 [98].

$$\mathrm{Imp}_{\mathrm{percentage(i)}} = \frac{\mathrm{Ocurr\_no(i)}}{\sum_{j=1}^{\mathrm{param\_no}} \mathrm{Occurr\_no(j)}} \tag{4}$$

As demonstrated in Fig. 7, researchers have focused on QoS factors: time by 24.2%, accuracy by 16.5%, scalability by 12.1%, Mean Square Error (MSE), Root Mean Square

Table 8 Categorization of recent papers by the hybrid approaches

References	Main idea	Evaluation type	Tool(s)	Advantage(s)	Disadvantage(s)
Manogaran and Lopez [90]	Climate change prediction based on a MapReduce algorithm	Real testbed	HBase Apache Hive Apache MapReduce framework CUSUM control chart	High prediction precision	High computational cost
Madan et al. [91]	Prefiguring the weather prediction by exploring dynamic statistical linear regression and supporting vector machine techniques	Data sets	Apache Hadoop	High prediction accuracy (Few errors)	Unable for concurrent prediction Time was not evaluated
Dhoot et al. [92]	Comparative analyzing for weather prediction with and without the need for Spark cluster	Real testbed	XGBoost Classifier	Low execution time	Low prediction accuracy for humidity data
Pandey et al. [8]	Integrating the Hadoop with ANFIS and FL methods for efficient weather prediction	Data sets	Apache Hadoop MATLAB Neural Network Fitting Tool Beautiful soup Python scripts	High prediction accuracy Low execution time	Low scalability
Bendre et al. [16]	Increasing the accuracy of the Weather forecasting based on MapReduce	Real testbed	Mahout Hbase Drill Storm-interactive	Low execution time High prediction accuracy High flexibility (performance with different datasets)	Precision, as mentioned in the title, was not attended to analyze the proposed algorithm
Dhamodharavadhani and Rathipriya [93]	Exponential smoothing rainfall prediction methods using MapReduce	Simulation	MATLAB MapReduce framework	High prediction accuracy Low execution time	Not good enough at initialization procedure
Namitha et al. [94]	Processing Big volume of weather Data based on MapReduce	Data sets	Apache Hadoop	Low execution time High scalability	Low reliability Low prediction accuracy
Liu et al. [95]	Forecasting weather using FP-tree algorithm based on MapReduce	Real testbed	Apache Hadoop	Low execution time High scalability	High Processing Overhead due to using Hadoop
Sahoo [96]	Forecasting weather using Incremental k-means clustering and Map Reduce	Real testbed	Apache Hadoop Apache pig	High reliability High security Low-cost commodity hardware High clustering performance	High processing Overhead due to using Hadoop
Fang et al. [97]	Classification and clustering Meteorological Data based on MapReduce	Real testbed	Apache Hadoop Mahout	Low execution time High scalability Low square error	High processing Overhead due to using Hadoop

Error (RMSE), precision, and reliability alike by 5.5%. It is also shown in Fig. 9 that they have used temperature 29.7%, wind 15.4%, precipitation 11%, humidity 12.1%, rainfall data 11%, and pressure 8.8%. Moreover, these results show that time, accuracy, scalability, MSE, RMSE, precision, and reliability factors are the most decisive QoS factors applied in the selected papers, and temperature, wind, precipitation, humidity, rainfall data, and pressure are the most used parameters in the reviewed papers. Figure 8 demonstrates the percentage of each QoS factor per approach. In the technique-based approaches, the accuracy with 19% percentage had the highest percentage, in the technology-based approaches, time with 40.9% had the highest percentage, and in the hybrid approaches, time with 25.9% had

Table 9 Comparison of the existing evaluation factors in the hybrid approaches

References	Temperature	Time	Wind	Accuracy	Scalability	Humidity	Precipitation	Rainfall data	Pressure	Mean square error	Reliability	Root mean square error	Precision
Manogaran and Lopez [90]	✓		✓			✓	✓	✓					✓
Madan et al. [91]	✓		✓	✓		✓			✓				
Dhoot et al. [92]	✓	✓		✓		✓							
Pandey et al. [8]	✓	✓		✓	✓	✓	✓			✓			
Bendre et al. [16]	✓	✓		✓		✓		✓					
Dhamodharavadhani and Rathipriya [93]		✓		✓				✓		✓			
Namitha et al. [94]	✓	✓		✓	✓			✓			✓	✓	✓
Liu et al. [95]	✓	✓	✓		✓	✓	✓		✓				
Sahoo [96]	✓		✓						✓		✓		
Fang et al. [97]	✓	✓	✓					✓	✓				

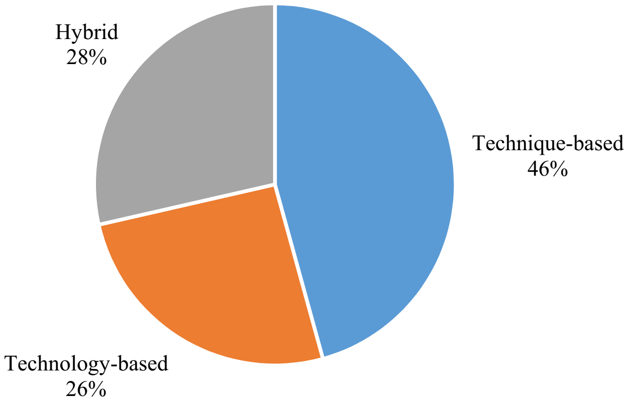

Fig. 6 Percentage of the reviewed approaches in big data analytics in weather forecasting

Table 10 The main advantages and disadvantages of the three approaches

Category	Advantages	Disadvantages
Technique-based	Better prediction accuracy Better execution time Better computational cost	Unacceptable reliability Unacceptable scalability Unacceptable simplicity
Technology-based	Better execution time Better scalability Better disk I/O	Unacceptable network latency Unacceptable processing overhead due to using Hadoop Unacceptable computational cost
Hybrid	Better prediction accuracy Better execution time Better scalability	Unacceptable reliability Unacceptable processing overhead due to using Hadoop Unacceptable computational cost

the highest percentage. Similarly, Fig. 10 represented the percentage of each weather parameter per approach. In the technique-based approaches, temperature with 28.9% and wind with 21.1% had the highest percentage; in the technology-based approaches, temperature with 41.2% and precipitation with 17.6% had the highest percentage, and in the hybrid approaches, temperature with 25% and humidity with 16.7% had the highest percentage. There were other QoS factors and weather parameters, including cost, Mean Absolute Error (MAE), solar, effectiveness, disk I/O, simplicity, visibility, sea level, efficiency, classification performance, reduction performance, prediction performance, clustering performance, recall, fault tolerance, security, and dew point.

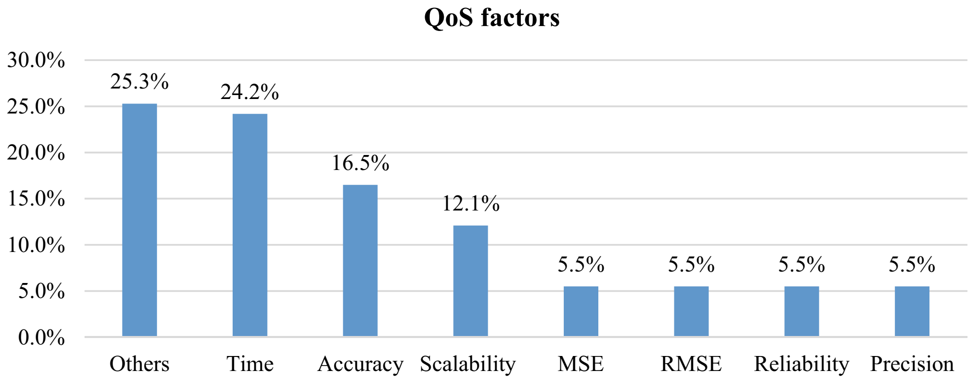

Fig. 7 Percentage of QoS factors investigated in the reviewed approaches

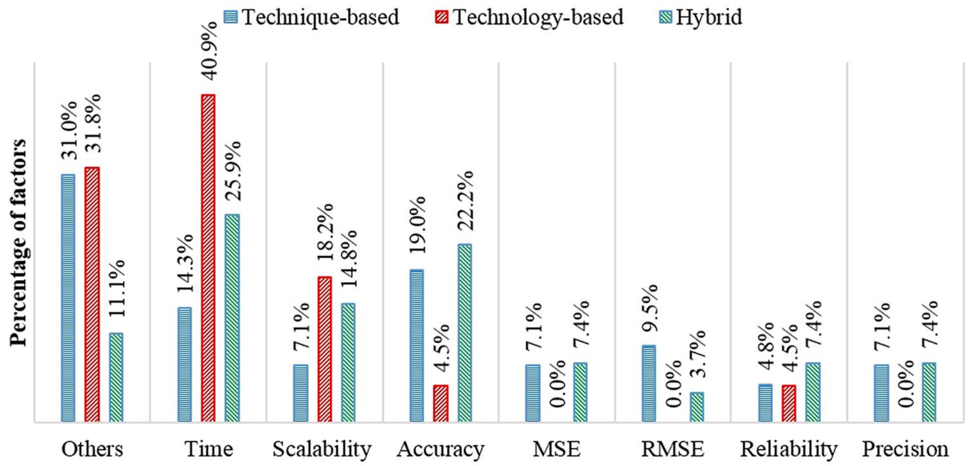

Fig. 8 Percentage of QoS factors in each category

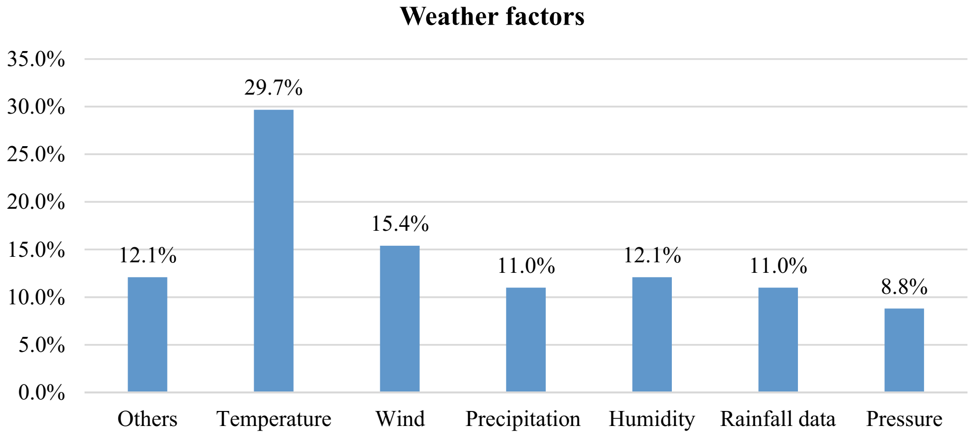

Fig. 9 Percentage of weather factors investigated in the reviewed approaches

- **RQ4**: What are the latest algorithms that help big data analytics in weather forecasting?

The used algorithms in three types of the mentioned approaches are divided into three classes. These classes and their statistic percentages, extracted from Tables 4, 6 and 8, are shown in Fig. 11. The greatest percentage of approaches used non-heuristic algorithms with 83%, meta-heuristic algorithms with 14%, and heuristic algorithms with 3%, respectively, the second and third ones. According to Tables 4, 6 and 8, the technique-based approaches applied meta-heuristic and non-heuristic algorithms; however, non-heuristic algorithms were applied most. The technology-based approaches used only non-heuristic algorithms. The hybrid approaches applied heuristic and non-heuristic, but non-heuristic algorithms were used most.

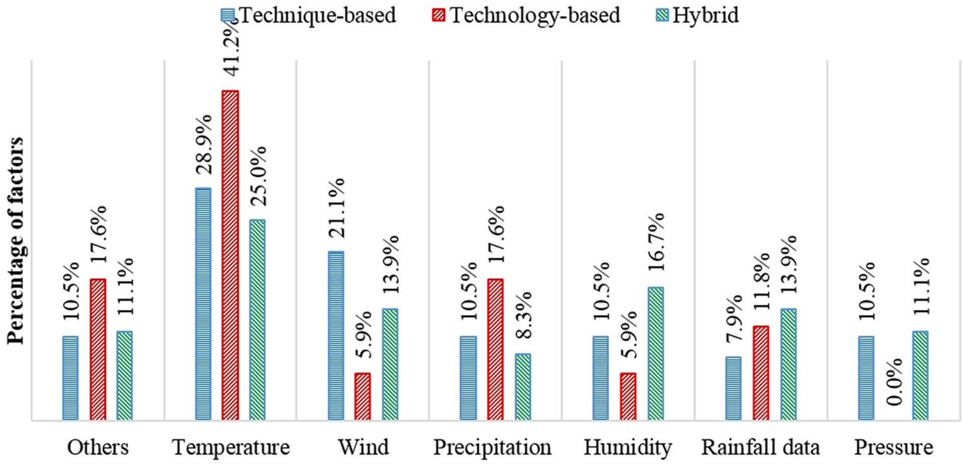

Fig. 10 Percentage of weather factors in each category

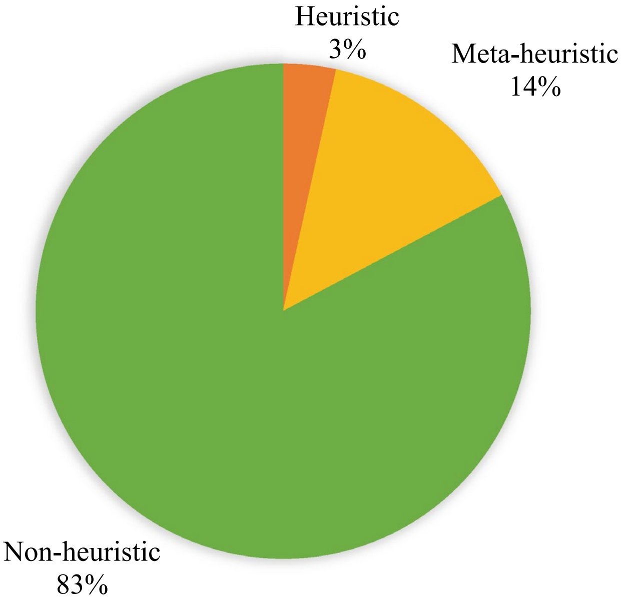

Fig. 11 Percentage of the algorithms supporting the big data analytics in weather forecasting

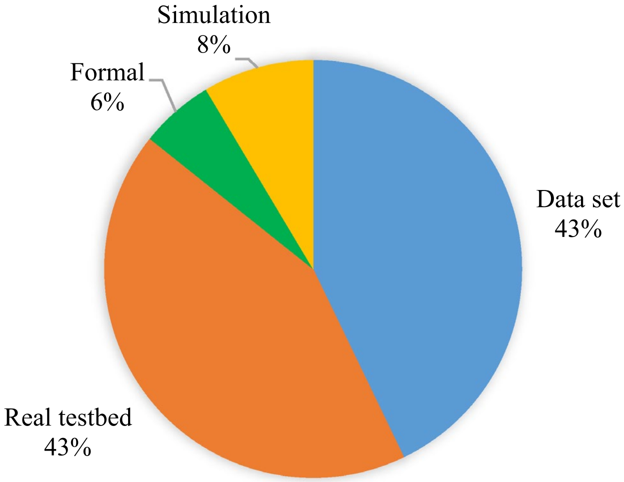

Fig. 12 Percentage of measurement environments used in the reviewed approaches

- **RQ5:** Which evaluation types are applied for evaluating the big data analytics in weather forecasting?

Based on Fig. 12, extracted from Tables 4, 6 and 8, we observed that 43% of the research papers used data sets to investigate their case studies. Also, 43% of approaches used a real testbed environment. In addition, the simulation environment had 8% of papers in this field. Lastly, it was observed that 6% of papers used formal evaluation. Based on the results summarized in Tables 4, 6 and 8, the technique-based approaches were implemented using data sets and formal environments or were simulated; data sets were used most. The technology-based approaches were only implemented in a real testbed. The hybrid approaches were implemented using data sets, real testbed environments, or were simulated; mostly real testbed was applied.

- **RQ6**: What are the utilized tools and frameworks for big data analytic approaches in weather forecasting, and how these are different?

The statistical percentage of the tools and frameworks used for this survey was extracted from Tables 4, 6 and 8, Apache Hadoop, with 38.6%, was used most in papers, MATLAB came next with 8.8%, Python by 7%, and Apache Spark was applied by 5.3%. Other tools and simulation environments were used by 28.1% in papers. As suggested by the results summarized in Tables 4, 6 and 8, the technology-based approaches employed Hadoop, Spark, Kafka, and Python. Hadoop and Spark were the most frequently used ones in this category. The technique-based approaches

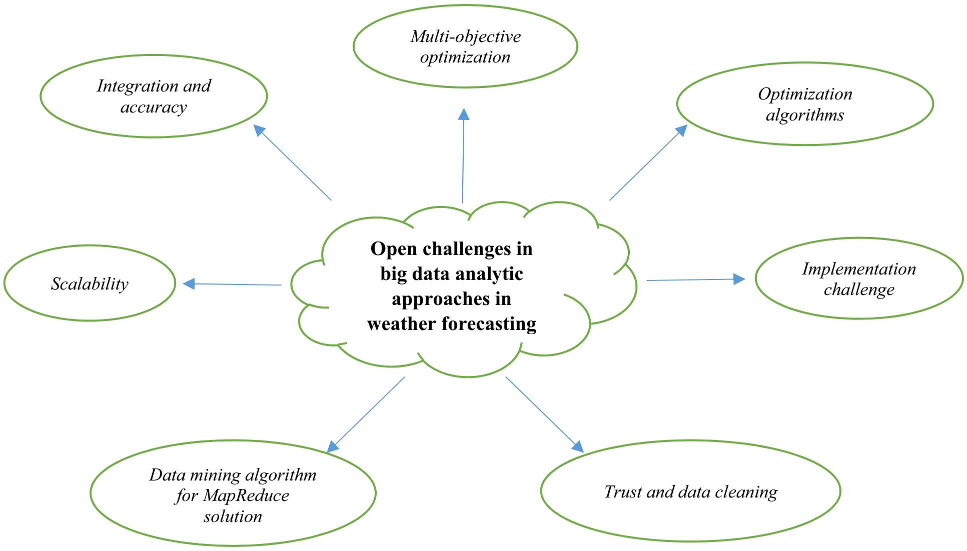
Multi-objective optimization
Integration and accuracy
Optimization algorithms
Open challenges in big data analytic approaches in weather forecasting
Scalability
Implementation challenge
Data mining algorithm for MapReduce solution
Trust and data cleaning

Fig. 13 Titles of open issues of big data analytics in weather forecasting

used Hadoop, MATLAB, and Python, but MATLAB was the highest frequently used one in this category. In the hybrid approaches, Hadoop, MATLAB, Mahout, and Python were used; Hadoop was the most frequently used tool in the hybrid category.

Both MATLAB and Python are common programming languages, but Python is more recent. There are differences between MATLAB and Python, and each one has its advantages and disadvantages as follows:

MATLAB is a programming language and interface that is generally used for technical computing. It has an easy-to-use visual computational environment. It also can be used for plotting functions. Python is an interactive high-level object-oriented programming language. It has an extensive standard library, and its implementation is very simple. XML processing, prototyping, and Web scripting in Python are rapid and easy. It is adaptable to the cloud environment. MATLAB is based on arrays, while Python is based on objects. Despite MATLAB, Python is free, open-source, and extensible software. It can be run on any servers and workstations. Its portability is higher than MATLAB. Everyone can modify the source code of Python. There are many packages that are developed for different task in Python. In MATLAB, developer-oriented add-ons should be validated and compiled before installation and usage. Testing algorithms in MATLAB can be immediately done without any compilation. It has high performance for technical computing.

On the other hand, both Hadoop and Spark are big data frameworks and are free. There are also differences between Hadoop and Spark, and each one has its advantages and disadvantages as follows:

Hadoop is based on batch processing, whereas Spark is based on real-time data processing. The latency of computing in Hadoop is higher than Spark. Spark is designed for fast computing. It has an interactive mode, but Hadoop does not have interactive data processing. Fault tolerance in Hadoop is achieved through replication, but Spark uses various data storage models and minimizes network I/O to achieve fault tolerance. The security feature of Hadoop is better than Spark. In Spark, everything is run in memory, and more RAM is required, but Hadoop is disk-bound and more systems are required to distribute the disk I/O.

7 Open Issues and Future Trends

Essentially, the accuracy of weather forecasts is important for meteorological service users. Also, the trust and response time of big data techniques and technologies are impressive for weather forecasters. In addition, some important open issues on big data analytics in weather forecasting that were not referred to completely by the research studies are depicted in Fig. 13.

- **RQ7:** What are the open issues and future trends of big data analytics in weather forecasting?
- *Multi-objective optimization*: While reviewing papers, we observed that big data analytic parameters were not examined in weather forecasting approaches at the same time. For instance, accuracy was measured in most of the technique-based approaches, while such factors as time, reliability, scalability, etc., were neglected. As a result, offering a multi-objective method for creating a trade-off between parameters will be a begging open issue.
- *Optimization algorithms*: According to the literature, different technique-based approaches in big data analytics for weather forecasting, such as clustering, classification, regression, and dimension reduction, are regarded as a subset of NP-hard or NP-complete category from the

facet of complexity. Applying heuristic or meta-heuristic techniques is, therefore, the best remedy for them. While several papers looked at such algorithms as genetic algorithm and swarm intelligence algorithm, other algorithms and optimization techniques like firefly algorithm [99], ant colony optimization [100, 101], particle swarm optimization [102], memetic algorithm [103, 104], bacterial colony optimization [105], simulated annealing [106], artificial immune system [107], learning automata [108], bat algorithm [109, 110], glowworm swarm optimization [111], and artificial bee colony [101, 112] were not appraised in the analyzed papers. Therefore, the mentioned optimization algorithms and techniques will be inspected as interesting future works.

- *Implementation challenge*: Big data analytic approaches conferred in weather forecasting should be operated in the real environment, while only 43% of surveyed papers were executed in a real environment, and most of them were tested with data sets, formal, and simulation environments. The mentioned approaches are required to be tested in a real environment for the performance test of the model. Therefore, real testbed implementation is very interesting. Also, we believe that the studied papers, with their data sets, formal or simulation evaluation, are also interesting to be executed in meteorological organizations.
- *Trust and data cleaning*: The challenges of noise, outliers, incomplete, and inconsistent data in traditional data management, according to the big data characteristics, will be increased for big data analytics. Data cleaning has been an important action in data management, especially in weather forecasting applications. Capturing and collecting weather data by different sensors and systems will make it easier to generate more incomplete and inconsistent data. Therefore, cleaning weather data is a necessity and yet a difficult process because knowing which values are trustworthy and which significant information should be to take out from such data is never easy. Hence, some efforts like developing combination classifiers on changing subsets of data [113] can be very challenging, so proposing more effective methods to be used in weather data cleaning systems is never easy.
- *Data mining algorithm for MapReduce solution*: Parallel computing is not applied to optimize most of the traditional data mining algorithms, so they are not specifically applicable to the big data area. Some researchers have tried to apply the traditional data mining algorithms to Hadoop-based platforms, such as parallel glowworm swarm optimization clustering algorithm on the basis of MapReduce [114] and parallel genetic algorithm in MapReduce [115]. As long as the transfer of data mining algorithms to Hadoop is unavoidable, applying other data mining algorithms to work on a MapReduce architecture is of great interest to future work.
- *Scalability*: Implementing the big data analytic in weather forecasting approaches in a large-scale environment is very necessary, but most researchers did not indicate if their approaches would run well in a large-scale environment. Even though the high scalability of the weather forecasting techniques is a meaningful tenet, based on the literature, most of the technique-based papers principally concentrated on approaches in small-scale methods. So, this issue is still an important, challenging topic.
- *Integration and accurate forecasts*: In addition to discovering more effective algorithms, integrating big data analytic technologies and techniques to forecast weather has proved to raise accuracy for the models. Numerical weather models integrated with machine learning algorithms or other methods seem to have better outcomes [79]. This integration opportunity can be an interdisciplinary problem and an open challenge.

8 Conclusion and Limitation

This review presented a systematic literature review of big data analytics in weather forecasting. At first, the concept of big data was introduced, and then, the concept of weather forecasting and the importance of big data in this era were studied. We discussed the research methodology and chose 185 papers published between 2014 and August 2020. Next and according to selection criteria, 35 papers were chosen. The selected studies were focused on three main categories: technique-based, technology-based, and hybrid approaches. The applied methods, advantages, and disadvantages of the discussed approaches were also specified. In addition, the QoS factors, weather parameters, types of algorithms, measurement environments, and modeling tools per paper were determined. According to the literature, 46% of the studied papers were technique-based, 28% of them belonged to the hybrid category, and 26% of papers were technology-based. Different types of algorithms that help big data analytics in weather forecasting were mentioned. As discussed in the paper, the technique-based approaches applied meta-heuristic and non-heuristic algorithms; however, non-heuristic algorithms were applied most. Technology-based approaches used only non-heuristic algorithms. Hybrid approaches applied heuristic and non-heuristic, but non-heuristic algorithms were used most.

In accordance with statistics, the QoS factors that the researchers focused on were time with 24.2%, accuracy with 16.5%, and scalability with 12.1%, MSE, RMSE, precision, and reliability alike with 5.5%. Also, weather parameters that they have used were temperature with 29.7%, wind with 15.4%, precipitation with 11%, humidity with 12.1%, rainfall data with 11%, and pressure with 8.8%. Moreover, most approaches used non-heuristic algorithms. We observed that 43% of the research papers used data sets to investigate their case studies. Also, 43% of approaches used a real testbed environment. In addition, the simulation environment had 8% of papers in this field. Lastly, it was observed that 6% of papers used formal evaluation. Based on the results, the technique-based approaches were implemented using data sets and formal environments or were simulated; data sets were used most. The technology-based approaches were only implemented in a real testbed. The hybrid approaches were implemented using data sets, real testbed environments, or were simulated; mostly real testbed was applied. We discussed the utilized modeling and simulation tools that the technology-based approaches employed Hadoop, Spark, Kafka, and Python. Hadoop and Spark were the most frequently used ones in this category. The technique-based approaches used Hadoop, MATLAB, and Python, but MATLAB was the highest frequently used one in this category. In the hybrid approaches, Hadoop, MATLAB, Mahout, and Python were used; Hadoop was the most frequently used tool in the hybrid category. Considering the open issues and future trends such as applying multi-objective optimization, implementation challenge, trust, and data cleaning, applying data mining algorithm for MapReduce, high scalability, integrating big data analytic technologies and techniques, and accurate forecasting can lead to more efficient big data analytic approaches for weather forecasting in the future.

We attempted to present a perfect systematic literature review of big data analytics in weather forecasting, and we encountered the following limitations:

- *Research sources*: Big data analytics in weather forecasting was introduced in several sources such as technical reports, academic publications, web pages, etc. To achieve the most qualified and trustworthy papers, only academic international conferences and journals were considered. In addition, nationally published conference papers and journals, non-English papers, and books were ignored.
- *Study and publication bias*: The reviewed papers were selected from several long-familiar databases such as IEEE, ScienceDirect, Springer, ACM, Wiley, Google Scholar, etc. Since these mentioned databases were proved to supply the most reliable and authentic papers, some related papers may have been omitted. Also, some related papers probably failed to be included through the selection processes declared in Sect. 4.
- *Research questions*: This paper was organized based on seven research queries; it is possible to regard more questions to write a more completed and detailed SLR.
- *Classification*: The reviewed papers were classified into three categories, including technique-based, technology-based, and hybrid approaches. Therefore, by increasing the number of published papers in the reviewed field, other classifications could be used by researchers.
- *Time scope*: Only the papers published from 2014 to August 2020 were selected.

We honestly wish the extracted results would assist researchers to reach more efficient approaches in big data analytics for weather forecasting in the future.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Xiao Z, Liu B, Liu H, Zhang D (2012) Progress in climate prediction and weather forecast operations in China. Adv Atmos Sci 29(5):943–957
2. Bengtsson L (1980) The weather forecast. Pure Appl Geophys 119(3):515–537
3. Kan L, Yu-Shu L (2005) A rough set based fuzzy neural network algorithm for weather prediction. In: 2005 International conference on machine learning and cybernetics, vol 3. pp 1888–1892
4. Kan L, Yu-Shu L (2002) Fuzzy case-based reasoning: weather prediction. In: Proceedings of the international conference on machine learning and cybernetics, vol 1. pp 107–110
5. Weiguo X (2010) The weather prediction method based on artificial immune system. In: 2010 International forum on information technology and applications, vol 2. pp 386–389
6. Haupt SE, Cowie J, Linden S, McCandless T, Kosovic B, Alessandrini S (2018) Machine learning for applied weather prediction. In: 2018 IEEE 14th international conference on e-science (e-Science). pp 276–277
7. Chung CYC, Kumar VR (1993) Knowledge acquisition using a neural network for a weather forecasting knowledge-based system. Neural Comput Appl 1(3):215–223
8. Pandey AK, Agrawal CP, Agrawal M (2017) A hadoop based weather prediction model for classification of weather data. In: 2017 Second international conference on electrical, computer and communication technologies (ICECCT). pp 1–5
9. Tsai C-W, Lai C-F, Chao H-C, Vasilakos AV (2015) Big data analytics: a survey. J Big Data 2(1):21
10. Rodríguez-Mazahua L, Rodríguez-Enríquez C-A, Sánchez-Cervantes JL, Cervantes J, García-Alcaraz JL, Alor-Hernández G (2016) A general perspective of big data: applications, tools, challenges and trends. J Supercomput 72(8):3073–3113
11. Talia D (2013) Clouds for scalable big data analytics. Computer 46(5):98–101
12. Selvaraj P, Marudappa P (2018) A survey of predictive analytics using big data with data mining. Int J Bioinf Res Appl 14:269
13. Sharma S, Mangat V (2015) Technology and trends to handle big data: survey. In: 2015 Fifth international conference on advanced computing and communication technologies. pp 266–271
14. Jain H, Jain R (2017) Big data in weather forecasting: applications and challenges. In: 2017 International conference on big data analytics and computational intelligence (ICBDAC). pp 138–142

15. Reddy PC, Babu AS (2017) Survey on weather prediction using big data analystics. In: 2017 Second international conference on electrical, computer and communication technologies (ICECCT). pp 1–6
16. Bendre MR, Thool RC, Thool VR (2015) Big data in precision agriculture: weather forecasting for future farming. In: 2015 1st international conference on next generation computing technologies (NGCT). pp 744–750
17. Mittal S, Sangwan OP (2019) Big data analytics using data mining techniques: a survey. In: Advanced informatics for computing research, Singapore. Springer Singapore, pp 264–273
18. Leu J-S, Su K-W, Chen C-T (2014) Ambient mesoscale weather forecasting system featuring mobile augmented reality. Multimed Tools Appl 72(2):1585–1609
19. Corne D, Dissanayake M, Peacock A, Galloway S, Owens E (2014) Accurate localized short term weather prediction for renewables planning. In: 2014 IEEE symposium on computational intelligence applications in smart grid (CIASG). pp 1–8
20. Roudier P et al (2014) The role of climate forecasts in smallholder agriculture: lessons from participatory research in two communities in Senegal. Clim Risk Mana 2:42–55
21. Li J, Xu L, Tang L, Wang S, Li L (2018) Big data in tourism research: a literature review. Tour Manag 68:301–323
22. Scott D, Lemieux C (2010) Weather and climate information for tourism. Procedia Environ Sci 1:146–183
23. Hazyuk I, Ghiaus C, Penhouet D (2012) Optimal temperature control of intermittently heated buildings using Model Predictive Control: Part I—Building modeling. Build Environ 51:379–387
24. Enríquez R, Jiménez MJ, Heras MdR (2016) Solar forecasting requirements for buildings MPC. Energy Procedia 91:1024–1032
25. Smith DA, Sherry L (2008) Decision support tool for predicting aircraft arrival rates from weather forecasts. In: 2008 Integrated communications, navigation and surveillance conference. pp 1–12
26. Zhang B, Tang L, Roemer M (2018) Probabilistic planning and risk evaluation based on ensemble weather forecasting. IEEE Trans Autom Sci Eng 15(2):556–566
27. Braman LM, van Aalst MK, Mason SJ, Suarez P, Ait-Chellouche Y, Tall A (2013) Climate forecasts in disaster management: red cross flood operations in West Africa, 2008. Disasters 37(1):144–164
28. Akhand MH (2003) Disaster management and cyclone warning system in Bangladesh. In: Zschau J, Küppers A (eds) Early warning systems for natural disaster reduction. Springer, Berlin, pp 49–64
29. Chen C, Duan S, Cai T, Liu B (2011) Online 24-h solar power forecasting based on weather type classification using artificial neural network. Sol Energy 85(11):2856–2870
30. Shi J, Lee W, Liu Y, Yang Y, Wang P (2012) Forecasting power output of photovoltaic systems based on weather classification and support vector machines. IEEE Trans Ind Appl 48(3):1064–1069
31. Lazos D, Sproul AB, Kay M (2014) Optimisation of energy management in commercial buildings with weather forecasting inputs: a review. Renew Sustain Energy Rev 39:587–603
32. Casas DM, González JÁT, Rodríguez JEA, Pet JV (2009) Using data-mining for short-term rainfall forecasting. In: Distributed computing, artificial intelligence, bioinformatics, soft computing, and ambient assisted living. Springer, Berlin, pp 487–490
33. Katal A, Wazid M, Goudar RH (2013) Big data: issues, challenges, tools and good practices. In: 2013 Sixth international conference on contemporary computing (IC3). pp 404–409
34. Elgendy N, Elragal A (2014) Big data analytics: a literature review paper. In: Advances in data mining. Applications and theoretical aspects. Springer, Cham, pp 214–227
35. Shadroo S, Rahmani A (2018) Systematic survey of big data and data mining in internet of things. Comput Netw 139:19–47
36. Bazzaz Abkenar S, Mahdipour E, Jameii SM, Haghi Kashani M (2021) A hybrid classification method for Twitter spam detection based on differential evolution and random forest. Concurr Comput Pract Exp. https://doi.org/10.1002/cpe.6381
37. Pathak AR, Pandey M, Rautaray S (2018) Construing the big data based on taxonomy, analytics and approaches. Iran J Comput Sci 1(4):237–259
38. Bazzaz Abkenar S, Haghi Kashani M, Mahdipour E, Jameii SM (2021) Big data analytics meets social media: a systematic review of techniques, open issues, and future directions. Telemat Inform 57:101517
39. Khezr SN, Navimipour NJ (2017) MapReduce and its applications, challenges, and architecture: a comprehensive review and directions for future research. J Grid Comput 15(3):295–321
40. Amer A-B, Amr M, Salah H (2016) A survey on MapReduce implementations. Int J Cloud Appl Comput IJCAC 6(1):59–87
41. Senger H et al (2016) BSP cost and scalability analysis for MapReduce operations. Concurr Comput Pract Exp 28(8):2503–2527
42. Lee D, Kim JW, Maeng S (2014) Large-scale incremental processing with MapReduce. Future Gener Comput Syst 36:66–79
43. Idris M et al (2015) Context-aware scheduling in MapReduce: a compact review. Concurr Comput Pract Exp 27(17):5332–5349
44. Karimi Y, Haghi Kashani M, Akbari M, Mahdipour E (2021) Leveraging big data in smart cities: a systematic review. J Concurr Comput Pract Exp. https://doi.org/10.1002/cpe.6379
45. Bakratsas M, Basaras P, Katsaros D, Tassiulas L (2018) Hadoop MapReduce performance on SSDs for analyzing social networks. Big Data Res 11:1–10
46. Shabestari F, Rahmani AM, Navimipour NJ, Jabbehdari S (2019) A taxonomy of software-based and hardware-based approaches for energy efficiency management in the Hadoop. J Netw Comput Appl 126:162–177
47. Patwardhan A, Verma AK, Kumar U (2016) A survey on predictive maintenance through big data. In: Current trends in reliability, availability, maintainability and safety. Springer, Cham, pp 437–445
48. Yang W, Liu X, Zhang L, Yang LT (2013) Big data real-time processing based on storm. In: 2013 12th IEEE international conference on trust, security and privacy in computing and communications. pp 1784–1787
49. Philip-Chen CL, Zhang C-Y (2014) Data-intensive applications challenges techniques and technologies: a survey on big data. Inf Sci 275:314–347
50. Lee J, Hong S, Lee J-H (2014) An efficient prediction for heavy rain from big weather data using genetic algorithm. In: Presented at the proceedings of the 8th international conference on ubiquitous information management and communication, Siem Reap, Cambodia
51. Sahasrabuddhe DV, Jamsandekar P (2015) Data structure for representation of big data of weather forecasting: a review. Int J Comput Sci Trends Technol IJCST 3(6):48–56
52. Priya SB A survey on weather forecasting to predict rainfall using big data analytics
53. Hassani H, Silva ES (2015) Forecasting with big data: a review. Ann Data Sci 2(1):5–19
54. Rao N (2017) Big data and climate smart agriculture-review of current status and implications for agricultural research and innovation in India. In: Proceedings Indian National Science Academy, Forthcoming
55. de Freitas Viscondi G, Alves-Souza SN (2019) A systematic literature review on big data for solar photovoltaic electricity generation forecasting. Sustain Energy Technol Assess 31:54–63

56. Vannitsem S et al (2021) Statistical postprocessing for weather forecasts: review, challenges, and avenues in a big data world. Bull Am Meteorol Soc 102(3):E681–E699
57. Cook DJ, Greengold NL, Ellrodt AG, Weingarten SR (1997) The relation between systematic reviews and practice guidelines. Ann Internal Med 127(3):210–216
58. Haghi Kashani M, Rahmani AM, Jafari Navimipour N (2020) Quality of service-aware approaches in fog computing. Int J Commun Syst 33(8):e4340
59. Rahimi M, Songhorabadi M, Haghi Kashani M (2020) Fog-based smart homes: a systematic review. J Netw Comput Appl 153:102531
60. Bazzaz Abkenar S, Haghi Kashani M, Akbari M, Mahdipour E (2020) Twitter spam detection: a systematic review. arXiv preprint arXiv:2011.14754.
61. Songhorabadi M, Rahimi M, Farid AMM, Kashani MH (2020) Fog computing approaches in smart cities: a state-of-the-art review. arXiv preprint arXiv:2011.14732
62. Kashani MH, Ahmadzadeh A, Mahdipour E (2020) Load balancing mechanisms in fog computing: a systematic review. arXiv preprint arXiv:2011.14706
63. Brereton P, Kitchenham BA, Budgen D, Turner M, Khalil M (2007) Lessons from applying the systematic literature review process within the software engineering domain. J Syst Softw 80(4):571–583
64. Sheikh Sofla M, Haghi Kashani M, Mahdipour E, Faghih Mirzaee R (2021) Towards effective offloading mechanisms in fog computing: a systematic survey. Multimed Tools Appl
65. Haghi Kashani M, Madanipour M, Nikravan M, Asghari P, Mahdipour E (2021) A systematic review of IoT in healthcare: applications, techniques, and trends. J Netw Comput Appl
66. Cheng Y, Zheng Z, Wang J, Yang L, Wan S (2019) Attribute reduction based on genetic algorithm for the coevolution of meteorological data in the industrial internet of things. Wirel Commun Mob Comput 2019:8
67. Cramer S, Kampouridis M, Freitas A (2016) A genetic decomposition algorithm for predicting rainfall within financial weather derivatives. In: Presented at the proceedings of the genetic and evolutionary computation conference 2016, Denver, Colorado, USA
68. Pooja SB, Siva-Balan RV, Anisha M, Muthukumaran MS, Jothikumar R (2020) Techniques Tanimoto correlated feature selection system and hybridization of clustering and boosting ensemble classification of remote sensed big data for weather forecasting. Comput Commun 151:266–274
69. Kvinge H, Farnell E, Kirby M, Peterson C (2018) Monitoring the shape of weather, soundscapes, and dynamical systems: a new statistic for dimension-driven data analysis on large datasets. In: 2018 IEEE international conference on big data (big data). pp 1045–1051
70. Buszta A, Mazurkiewicz J (2015) Climate changes prediction system based on weather big data visualisation. In: Theory and engineering of complex systems and dependability. Springer, Cham, pp 75–86
71. Rasel RI, Sultana N, Meesad P (2018) An application of data mining and machine learning for weather forecasting. In: Recent advances in information and communication technology 2017. Springer, Cham, pp 169–178
72. Mahmood MR, Patra RK, Raja R, Sinha GR (2019) A novel approach for weather prediction using forecasting analysis and data mining techniques. In: Innovations in electronics and communication engineering. Springer, Singapore, pp 479–489
73. Azimi R, Ghofrani M, Ghayekhloo M (2016) A hybrid wind power forecasting model based on data mining and wavelets analysis. Energy Convers Manag 127:208–225
74. Doreswamy IG, Manjunatha BR (2018) Multi-label classification of big NCDC weather data using deep learning model. In: Soft computing systems. Springer, Singapore, pp 232–241
75. Venkatachalapathy K, Kamaleshwar T, Sundaranarayana D, Prakash VO (2016) An effective framework with N-client transfer dataset for weather prediction using data mining techniques. In: Presented at the proceedings of the international conference on informatics and analytics, Pondicherry, India
76. Choi C, Kim J, Kim J, Kim D, Bae Y, Kim HS (2018) Development of heavy rain damage prediction model using machine learning based on big data. Adv Meteorol 2018:11
77. Hubig N, Fengler P, Züfle A, Yang R, Günnemann S (2017) Detection and prediction of natural hazards using large-scale environmental data. In: Advances in spatial and temporal databases. Springer, Cham, pp 300–316
78. Yonekura K, Hattori H, Suzuki T (2018) Short-term local weather forecast using dense weather station by deep neural network. In: 2018 IEEE international conference on big data (big data). pp 1683–1690
79. Xu Q et al (2015) A short-term wind power forecasting approach with adjustment of numerical weather prediction input by data mining. IEEE Trans Sustain Energy 6(4):1283–1291
80. Jiang P, Dong Q (2015) A new hybrid model based on an intelligent optimization algorithm and a data denoising method to make wind speed predication. Math Probl Eng 2015:16
81. More PD, Nandgave S, Kadam M (2020) Weather data analytics using hadoop with map-reduce. In: ICCCE 2019. Springer, Singapore, pp 189–196
82. Wu H (2017) Big data management the mass weather logs. In: Smart computing and communication. Springer, Cham, pp 122–132
83. Ismail KA, Majid MA, Zain JM, Bakar NAA (2016) Big data prediction framework for weather temperature based on MapReduce algorithm. In: 2016 IEEE conference on open systems (ICOS). pp 13–17
84. Abdullahi AU, Ahmad R, Zakaria NM (2016) Big data: performance profiling of meteorological and oceanographic data on hive. In: 2016 3rd international conference on computer and information sciences (ICCOINS). pp 203–208
85. Oury DTM, Singh A (2018) Data analysis of weather data using hadoop technology. In: Smart computing and informatics. Springer, Singapore, pp 723–730
86. Manogaran G, Lopez D, Chilamkurti N (2018) In-mapper combiner based MapReduce algorithm for processing of big climate data. Future Gener Comput Syst 86:433–445
87. Jayanthi D, Sumathi G (2017) Weather data analysis using spark—an in-memory computing framework. In: 2017 Innovations in power and advanced computing technologies (i-PACT). pp 1–5
88. Palamuttam R et al (2015) SciSpark: Applying in-memory distributed computing to weather event detection and tracking. In: 2015 IEEE International conference on big data (big data). pp 2020–2026
89. Hassaan M, Elghandour I (2016) A real-time big data analysis framework on a CPU/GPU heterogeneous cluster: a meteorological application case study. In: 2016 IEEE/ACM 3rd international conference on big data computing applications and technologies (BDCAT). pp 168–177
90. Manogaran G, Lopez D (2018) Spatial cumulative sum algorithm with big data analytics for climate change detection. Comput Electr Eng 65:207–221
91. Madan S, Kumar P, Rawat S, Choudhury T (2018) Analysis of weather prediction using machine learning & big data. In: 2018 International conference on advances in computing and communication engineering (ICACCE). pp 259–264

92. Dhoot R, Agrawal S, Kumar MS (2019) Implementation and analysis of arima model and kalman filter for weather forcasting in spark computing environment. In: 2019 3rd international conference on computing and communications technologies (ICCCT). pp 105–112
93. Dhamodharavadhani S, Rathipriya R (2019) Region-wise rainfall prediction using mapreduce-based exponential smoothing techniques. In: Advances in big data and cloud computing. Springer, Singapore, pp 229–239
94. Namitha K, Jayapriya A, Kumar GS (2015) Rainfall prediction using artificial neural network on map-reduce framework. In: Presented at the proceedings of the third international symposium on women in computing and informatics, Kochi, India
95. Liu L, Lv J, Ma Z, Wan J, Jingjing M (2015) Toward the association rules of meteorological data mining based on cloud computing. In: Proceedings of the second international conference on mechatronics and automatic control. Springer, Cham, pp 1051–1059
96. Sahoo S (2017) A parallel forecasting approach using incremental K-means clustering technique. In: Computational intelligence in data mining. Springer, Singapore, pp 165–172
97. Fang W, Sheng VS, Wen X, Pan W (2014) Meteorological data analysis using MapReduce. Sci World J 2014:10
98. Hamzei M, Navimipour NJ (2018) Toward efficient service composition techniques in the internet of things. IEEE Internet Things J 5(5):3774–3787
99. Kumar V, Kumar D (2020) A systematic review on firefly algorithm: past, present, and future. Arch Comput Methods Eng 28(4):3269–3291
100. Nikravan M, Kashani MH (2007) Parallel min–max ant colony system (MMAS) for dynamic process scheduling in distributed operating systems considering load balancing. In: Proceedings of the 21st ECMS international conference on high performance computing & simulation (HPCS), Prague, Czech Republic
101. Kashani MH, Sarvizadeh R (2011) A novel method for task scheduling in distributed systems using max–min ant colony optimization. In: 2011 3rd international conference on advanced computer control (ICACC). IEEE, pp 422–426
102. Kashani MH, Zarrabi H, Javadzadeh G (2017) A new metaheuristic approach to task assignment problem in distributed systems. In: 2017 IEEE 4th international conference on knowledge-based engineering and innovation (KBEI). IEEE, pp 0673–0677
103. Kashani MH, Sarvizadeh R, Jameii M (2012) A new distributed systems scheduling algorithm: a swarm intelligence approach. In: Fourth international conference on machine vision (ICMV 2011): computer vision and image analysis; pattern recognition and basic technologies. International Society for Optics and Photonics
104. Kashani MH, Jahanshahi M (2009) A new method based on memetic algorithm for task scheduling in distributed systems. Int J Simul Syst Sci Technol 10
105. Niu B, Wang H (2012) Bacterial colony optimization. Discrete Dyn Nat Soc 2012:28
106. Kashani MH, Jahanshahi M (2009) Using simulated annealing for task scheduling in distributed systems. In: 2009 International conference on computational intelligence, modelling and simulation. pp 265–269
107. Dasgupta D, Ji Z, Gonzalez F (2003) Artificial immune system (AIS) research in the last five years. In: The 2003 congress on evolutionary computation, 2003. CEC '03., vol 1. pp 123–130
108. Jameii SM, Kashani MH, Karimi R (2015) LASPEA: Learning automata-based strength pareto evolutionary algorithm for multi-objective optimization. Int J Comput Sci Telecommun 6(9):14–19
109. Yang X-S (2010) A new metaheuristic bat-inspired algorithm. In: Nature inspired cooperative strategies for optimization (NICSO 2010). Springer, pp 65–74
110. Yang X-S. Bat algorithm for multi-objective optimisation. arXiv e-prints, Accessed 01 Mar 2012. arXiv:1203.6571Y
111. Krishnanand KN, Ghose D (2009) Glowworm swarm optimization for simultaneous capture of multiple local optima of multimodal functions. Swarm Intell 3(2):87–124
112. Sarvizadeh R, Kashani MH, Zakeri FS, Jameii SM (2012) A novel bee colony approach to distributed systems scheduling. Int J Comput Appl 42(10):1–6
113. Saneja B, Rani R (2018) A hybrid approach for outlier detection in weather sensor data. In: 2018 IEEE 8th international advance computing conference (IACC). pp 321–326
114. Al-Madi N, Aljarah I, Ludwig S (2014) Parallel Glowworm Swarm Optimization Clustering Algorithm based on MapReduce
115. El-Alfy E-SM, Alshammari MA (2016) Towards scalable rough set based attribute subset selection for intrusion detection using parallel genetic algorithm in MapReduce. Simul Model Pract Theory 64(13):18–29