Chapter 4
Rainfall Prediction Using Machine Learning Models: Literature Survey

Eslam A. Hussein, Mehrdad Ghaziasgar, Christopher Thron, Mattia Vaccari, and Yahlieel Jafta

4.1 Introduction

Natural processes on Earth can be classified into several categories, including hydrological processes like storm waves and groundwater; biological processes like forest growth; atmospheric processes like thunderstorms and rainfall; human processes like urban development; and geological processes like earthquakes. The field of physical geography seeks to investigate the distribution of the different features/parameters that describe the landscape and functioning of the Earth by analyzing the processes that shape it. These features/parameters have been referred to as geophysical parameters in the literature (Karimi, 2014).

Rainfall is a key geophysical parameter that is essential for many applications in water resource management, especially in the agriculture sector. Predicting rainfall can help managers in various sectors to make decisions regarding a range of important activities such as crop planting, traffic control, the operation of sewer systems, and managing disasters like droughts and floods (Htike and Khalifa, 2010). A number of countries such as Malaysia and India depend on the agriculture sector as a major contributor to the economy (Htike and Khalifa, 2010; Parmar et al., 2017) and as a source of food security. Hence, an accurate prediction of rainfall is needed

E. A. Hussein (✉) · M. Ghaziasgar · Y. Jafta
Department of Computer Science, University of the Western Cape, Cape Town, South Africa
e-mail: ehussein@uwc.ac.za; mghaziasgar@uwc.ac.za; 2858132@myuwc.ac.za

C. Thron
Department of Science and Mathematics, Texas A&M University-Central, Killeen, TX, USA
e-mail: thron@tamuct.edu

M. Vaccari
Department of Physics and Astronomy, University of the Western Cape, Cape Town, South Africa
e-mail: mvaccari@uwc.ac.za

M. Alloghani et al. (eds.), *Artificial Intelligence for Data Science in Theory and Practice*, Studies in Computational Intelligence 1006,
https://doi.org/10.1007/978-3-030-92245-0_4

to make better future decisions to help manage activities such as the ones mentioned above.

Rainfall is considered to be one of the most complicated parameters to forecast in the hydrological cycle (Htike and Khalifa, 2010; Hung et al., 2009; Nasseri et al., 2008). This is due to the dynamic nature of environmental factors and random variations, both spatially and temporally, in these factors (Htike and Khalifa, 2010). Therefore, to address random variations in rainfall, several machine learning (ML) tools including artificial neural networks (ANN), *k*-nearest neighbours (KNNs), decision trees (DT), etc. are used in the literature to learn patterns in the data to forecast rainfall. In this chapter, a review of past work in the area of rainfall prediction using ML models is carried out.

A number of related review papers exist as follows. The authors in Mosavi et al. (2018) focused on reviewing studies that use ML for flood prediction, which closely resembles rainfall prediction. The authors in Shi and Yeung (2018) focused on the use of ML for generic spatio-temporal sequence forecasting. Finally, the authors in Parmar et al. (2017) conducted a survey on the use of ML for rainfall prediction: however, the study was limited to rainfall prediction in India.

This chapter serves as an addition to the field by surveying recent relevant studies focusing on the use of ML in rainfall prediction in a variety of geographic locations from 2016–2020. After detailing the methods used to forecast rainfall, one of the important contributions of this chapter is to demonstrate various pitfalls that lead to an overestimation in model performance of the ML models in various papers. This in turn leads to unrealistic hype and expectations surrounding ML in the current literature. It also leads to an unrealistic understanding of the advancements in, and gains by, ML research in this field. It is therefore important to clearly state and demonstrate these pitfalls in order to help researchers avoid them.

The rest of this review is organized as follows: Section 4.2 discusses the methodology used to survey and review the literature which defines the discussion framework used in all subsequent sections; Sect. 4.3 describes the data sets used; Sect. 4.4 provides a description of the output objective in the various papers; Sects. 4.5–4.7 describe the input features used, common methods of pre-processing and the ML models used; Sect. 4.8 summarizes the results obtained in various studies; and Sect. 4.9 then provides a discussion of the procedures used, specifically pointing out the pitfalls mentioned before towards obtaining over-estimated and unrealistic results. The section that follows concludes the paper.

4.2 Methodology

This chapter carries out an in-depth review of relevant literature to reveal the different practices authors take to predict rainfall. The review covers several aspects which relate to the input into, output from, and methods used in the various systems devised in the literature for this purpose. The review specifically focuses on studies that use supervised learning for both regression and classification problems.

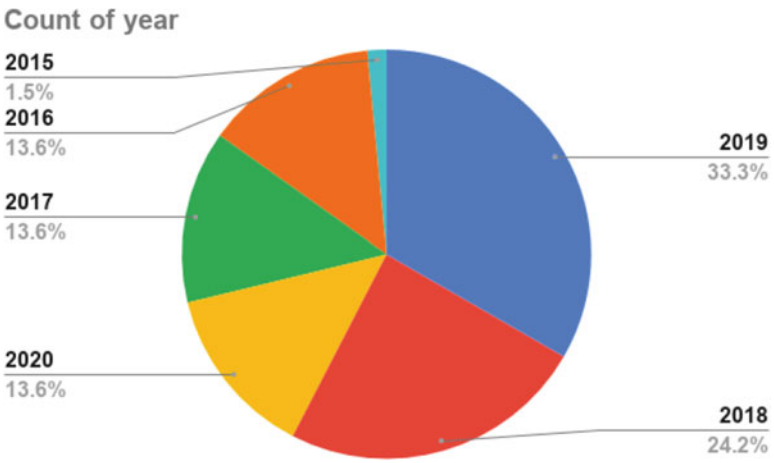

Fig. 4.1 Pie chart showing proportions by publication year for papers in this review

Google scholar was used to collect papers from 2016 to 2020, with the following key words: ("machine learning" OR "deep learning") AND ("precipitation prediction" OR "rainfall prediction" OR "precipitation nowcasting"). Almost 1240 results were obtained, and of these only supervised rainfall prediction papers that used meteorological data from, e.g., radar, satellites, and stations were selected, while papers that used data from normal cameras, e.g., photographs were excluded. Even though this review focuses on the prediction of rainfall, the methods used to achieve this can be extended and applied to other geophysical parameters like temperature and wind. Hence, the conclusions and discussions of this chapter can be adapted to other parameters.

The total number of reviewed papers are 66, which are a combination of conferences and journal papers published from 2016–2020, except for one paper (Shi et al., 2015) which was published in 2015 and is a seminal work in this field. Figure 4.1 shows the reviewed studies per year. Tables which summaries the reviewed paper can be found in Appendices 1 and 2.

Figure 4.2 shows the generic structure of supervised ML models. This structure was used as a guideline to construct a set of questions used to systematically categorize and analyze the 66 papers. The questions are as follows:

1. What data sets are used and where are they sourced?
2. What is the output objective in the various papers in terms of what the goal of prediction/forecasting?
3. What input features are extracted from the data set(s) to be used to achieve the output objective?
4. What pre-processing methods are used prior to classification/regression?
5. What ML models are used to achieve classification/regression towards the output objective?

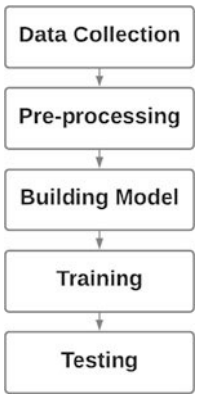

Fig. 4.2 Basic flow for building machine learning (ML) models (Mosavi et al., 2018)

6. What results were obtained from the above-mentioned steps, and how were they reported?

These questions provide the framework for the rest of this paper. Sections 4.3–4.8 address questions 1–6 in sequence. Section 4.9 discusses the findings in the previous six sections, and Sect. 4.10 provides conclusions.

4.3 Data Sets

This section provides a breakdown of the data sets used in the 66 studies surveyed, based on the sources of the data sets, availability, and geographical locations where the data sets were collected.

Figure 4.3 (left) provides a breakdown of the studies based on the sources/availability of the data sets used in those studies. About 75% of the studies used private data, sourced from meteorological stations of their prospective countries (Peng et al., 2019; Pham et al., 2019; Zhang et al., 2020, 2018; Zainudin et al., 2016; Singh and Kumar, 2019; Oswal, 2019; Balamurugan and Manojkumar, 2021; Du et al., 2017; Manandhar et al., 2019; Kashiwao et al., 2017; Shi et al., 2015; Singh et al., 2017; Jing et al., 2019; Ayzelet al., 2019; Chen et al., 2020; Tran and Song, 2019a; Shi et al., 2017; Wang et al., 2017; Tran and Song, 2019b; Shi et al., 2017; Du et al., 2018; Chattopadhyay et al., 2020; Zhuang and Ding, 2016; Du et al., 2019; Dash et al. , 2018; Cristian, 2018; Lakshmaiah et al., 2016; Ramsundram et al., 2016; Sulaiman and Wahab, 2018; Amiri et al., 2016; Abbot and Marohasy, 2016; Sardeshpande and Thool, 2019; Shenify et al., 2016; Banadkooki et al. , 2019; Mehr et al., 2019; Beheshti et al., 2016; Nourani et al., 2019; Abbot and Marohasy, 2017; Kumar et al., 2019; Haidar and Verma, 2018; Duong et al., 2018; Xu et al., 2020; Aswin et al., 2018; Chhetri et al., 2020; Bojang et al., 2020; Canchala et al., 2020; Mehdizadeh et al., 2018; Weesakul et al., 2018; Gao et al., 2019; Mishra and Kushwaha, 2019). Most of these data sets are not readily available for use. Only

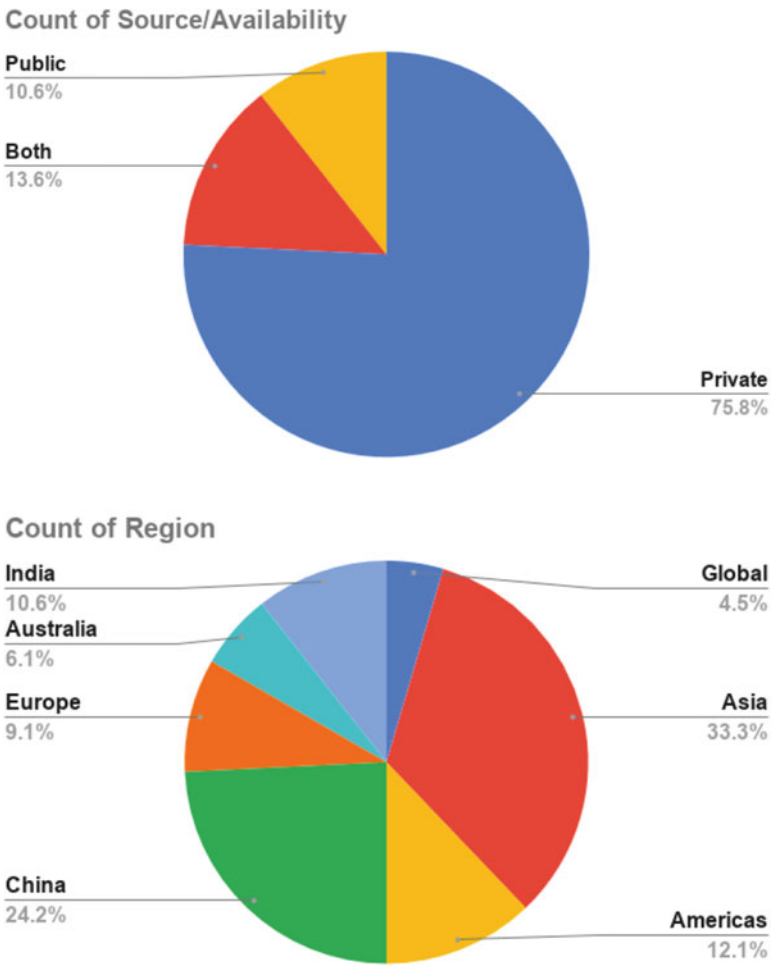

Fig. 4.3 Pie chart of the percentage of data sets in this survey in terms of source/availability (top) and geographical region (bottom)

10% of the studies use data sourced from freely available sources such as Kaggle (www.kaggle.com), and the National Oceanic and Atmospheric Administration (NOAA) (Sato et al., 2018; Castro et al., 2020; Pan et al., 2019; Zhan et al., 2019; Patel et al., 2018; Damavandi and Shah, 2019; Aswin et al., 2018). The remaining 13% of studies in this review use data from both private and publicly available sources (Valencia-Payan and Corrales, 2018; Yu et al., 2017; Diez-Sierra and del Jesus, 2020; Chen et al., 2016; Huang et al., 2017; Boonyuen et al., 2018, 2019; Lee et al., 2018; Aguasca-Colomo et al., 2019).

Figure 4.3 (right) summarizes the geographical regions included in this review. The continent of Asia accounts for around 68% of all studies (Pham et al., 2019; Yu

et al., 2017; Zainudin et al., 2016; Manandhar et al., 2019; Kashiwao et al., 2017; Shi et al., 2015; Sato et al., 2018; Shi et al., 2017; Boonyuen et al., 2018, 2019; Sulaiman and Wahab, 2018; Amiri et al., 2016; Banadkooki et al. , 2019; Mehr et al., 2019; Damavandi and Shah, 2019; Lee et al., 2018; Beheshti et al., 2016; Duong et al., 2018; Chhetri et al., 2020; Bojang et al., 2020; Mehdizadeh et al., 2018; Weesakul et al., 2018; Zhang et al., 2018; Chen et al., 2016; Du et al., 2017; Huang et al., 2017; Jing et al., 2019; Chen et al., 2020; Tran and Song, 2019a; Shi et al., 2017; Castro et al., 2020; Wang et al., 2017; Tran and Song, 2019b; Du et al., 2018, 2019; Xu et al., 2020; Gao et al., 2019; Balamurugan and Manojkumar, 2021; Dash et al. , 2018; Lakshmaiah et al., 2016; Ramsundram et al., 2016; Sardeshpande and Thool, 2019; Kumar et al., 2019; Mishra and Kushwaha, 2019). Of these, studies that focus on China and India make up almost one quarter and one tenth respectively of all studies in this review. The remaining Asian studies focus on countries such as Iran, South Korea and Japan.

The rest of the chart is distributed as follows: the Americas make up 12.1% of studies (Valencia-Payan and Corrales, 2018; Singh and Kumar, 2019; Singh et al., 2017; Pan et al., 2019; Zhan et al., 2019; Chattopadhyay et al., 2020; Zhuang and Ding, 2016; Canchala et al., 2020); Europe accounts for 9.1% (Diez-Sierra and del Jesus, 2020; Ayzelet al., 2019; Cristian, 2018; Shenify et al., 2016; Nourani et al., 2019; Aguasca-Colomo et al., 2019); Australia comprises 6.1%; (Oswal, 2019; Abbot and Marohasy, 2016, 2017; Haidar and Verma, 2018), and the remaining 4.5% either involve multiple regions or involve the use of the whole global map (Peng et al., 2019; Patel et al., 2018; Aswin et al., 2018).

4.4 Output Objectives

The output objectives of rainfall forecasting studies can be analyzed in terms of three factors: the forecasting time frame of the output; whether the output is continuous or discrete; and the dimensionality of the output. The forecasting time frame of the output specifies the time span of the forecast made, i.e., hourly, daily, monthly, etc. The output can also be discrete (e.g., classification into "Rain"/"No Rain" classes), or continuous (e.g., predicting the quantity of rain), or both. Finally, the output can be 1-dimensional (1D) in the form of a single number or label representing a rainfall measure or category, or 2-dimensional (2D) in the form of a geospatial map of rainfall measures or categories on a grid of the geographical location under study.

In terms of the forecasting time frame, the studies can be broken down into those that make long-term predictions and those that focus on making short-term predictions. In this review, long-term prediction is defined as predictions of one month up to a year ahead, while short-term prediction can be a few minutes ahead (e.g., 5–15 minutes), up to one or more days ahead. Figure 4.4 (left) shows the distribution of papers' forecasting time frames. Of the 66 reviewed papers, 30 papers (45%) make long-term predictions, the majority of which focus on

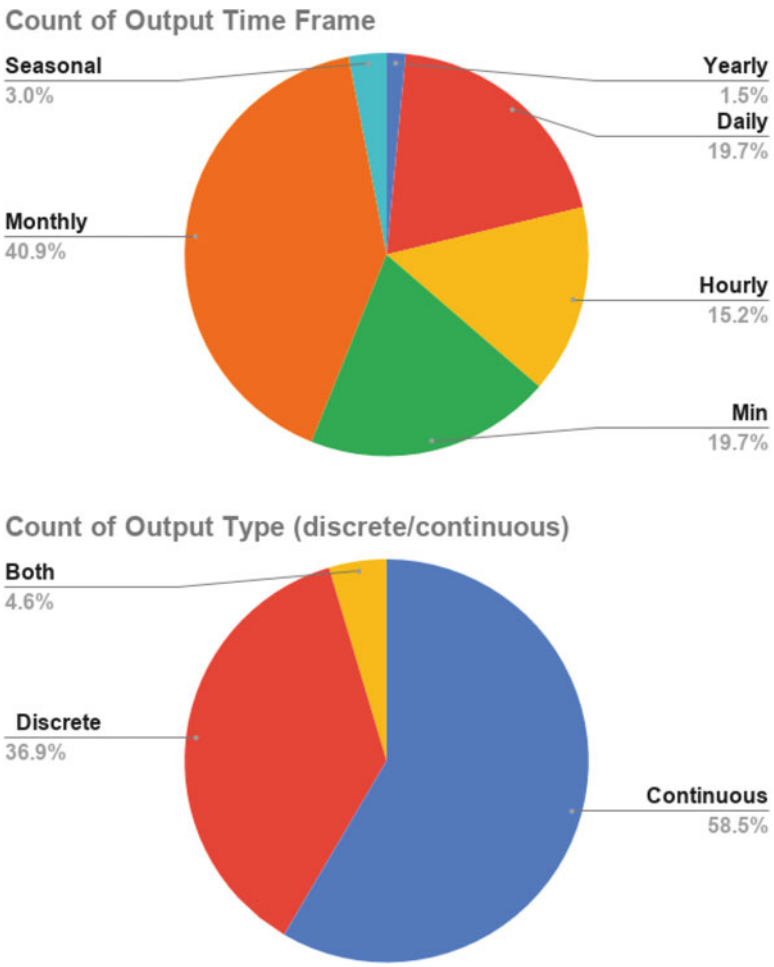

Fig. 4.4 Pie chart of the percentage of data sets in this survey in terms of forecasting time frame (top) and the discrete (classification)/continuous (regression) nature of the prediction output (bottom)

monthly forecasting (Cristian, 2018; Lakshmaiah et al., 2016; Ramsundram et al., 2016; Sulaiman and Wahab, 2018; Amiri et al., 2016; Abbot and Marohasy, 2016; Sardeshpande and Thool, 2019; Shenify et al., 2016; Banadkooki et al. , 2019; Mehr et al., 2019; Damavandi and Shah, 2019; Lee et al., 2018; Beheshti et al., 2016; Nourani et al., 2019; Abbot and Marohasy, 2017; Kumar et al., 2019; Haidar and Verma, 2018; Duong et al., 2018; Xu et al., 2020; Aswin et al., 2018; Canchala et al., 2020; Bojang et al., 2020; Chhetri et al., 2020; Mehdizadeh et al., 2018; Weesakul et al., 2018; Mishra and Kushwaha, 2019; Aguasca-Colomo et al., 2019). Only two studies focus on seasonal forecasting (Du et al., 2019; Dash et al. , 2018),while a

single study aims towards yearly forecasting (Gao et al., 2019). As for studies that focus on short-term prediction, these are broken down nearly evenly between daily (Peng et al., 2019; Pham et al., 2019; Zainudin et al., 2016; Diez-Sierra and del Jesus, 2020; Singh and Kumar, 2019; Oswal, 2019; Balamurugan and Manojkumar, 2021; Huang et al., 2017; Castro et al., 2020; Pan et al., 2019; Zhuang and Ding, 2016; Boonyuen et al., 2018, 2019), hourly (Valencia-Payan and Corrales, 2018; Zhang et al., 2020, 2018; Yu et al., 2017; Chen et al., 2016; Du et al., 2017, 2018; Zhan et al., 2019; Chattopadhyay et al., 2020; Patel et al., 2018), and one or more minutes ahead (Manandhar et al., 2019; Kashiwao et al., 2017; Shi et al., 2015; Singh et al., 2017; Jing et al., 2019; Sato et al., 2018; Ayzelet al., 2019; Chen et al., 2020; Tran and Song, 2019a; Shi et al., 2017; Wang et al., 2017; Tran and Song, 2019b; Shi et al., 2017).

In terms of the type of output, i.e., discrete (classification) or continuous (regression), Fig. 4.4 (right) shows the distribution between the different output types. The majority carried out regressions to obtain continuous output (Du et al., 2019; Dash et al. , 2018; Cristian, 2018; Lakshmaiah et al., 2016; Ramsundram et al., 2016; Sulaiman and Wahab, 2018; Amiri et al., 2016; Abbot and Marohasy, 2016; Sardeshpande and Thool, 2019; Shenify et al., 2016; Banadkooki et al. , 2019; Mehr et al., 2019; Damavandi and Shah, 2019; Lee et al., 2018; Beheshti et al., 2016; Nourani et al., 2019; Abbot and Marohasy, 2017; Kumar et al., 2019; Haidar and Verma, 2018; Duong et al., 2018; Xu et al., 2020; Aswin et al., 2018; Canchala et al., 2020; Bojang et al., 2020; Chhetri et al., 2020; Mehdizadeh et al., 2018; Weesakul et al., 2018; Peng et al., 2019; Pham et al., 2019; Valencia-Payan and Corrales, 2018; Zhang et al., 2020, 2018; Yu et al., 2017; Tran and Song, 2019a; Wang et al., 2017; Shi et al., 2017; Du et al., 2018), while slightly more than one third carried out classification into discrete classes (Zainudin et al., 2016; Diez-Sierra and del Jesus, 2020; Singh and Kumar, 2019; Oswal, 2019; Balamurugan and Manojkumar, 2021; Chen et al., 2016; Du et al., 2017; Manandhar et al., 2019; Kashiwao et al., 2017; Huang et al., 2017; Shi et al., 2015; Singh et al., 2017; Jing et al., 2019; Sato et al., 2018; Chen et al., 2020; Shi et al., 2017; Chattopadhyay et al., 2020; Patel et al., 2018; Zhuang and Ding, 2016; Boonyuen et al., 2018, 2019; Gao et al., 2019; Mishra and Kushwaha, 2019; Aguasca-Colomo et al., 2019). Only 3 studies applied both classification and regression (Ayzelet al., 2019; Tran and Song, 2019b; Zhan et al., 2019).

For studies that applied classification,mostly carried out binary classification (Zainudin et al., 2016; Diez-Sierra and del Jesus, 2020; Singh and Kumar, 2019; Oswal, 2019; Balamurugan and Manojkumar, 2021; Chen et al., 2016; Du et al., 2017; Manandhar et al., 2019; Kashiwao et al., 2017; Huang et al., 2017; Shi et al., 2015; Singh et al., 2017; Jing et al., 2019; Sato et al., 2018; Chen et al., 2020; Shi et al., 2017; Patel et al., 2018; Zhuang and Ding, 2016; Boonyuen et al., 2018), with the majority of these classified into “Rain"/“No Rain" classes. Relatively fewer studies aim towards carrying out classification into multiple classes (Huang et al., 2017; Chattopadhyay et al., 2020; Boonyuen et al., 2019; Gao et al., 2019; Mishra and Kushwaha, 2019), varying from three to five classes.

Finally, for the dimensionality of the output, 54 out of 66 studies produce 1D output (Du et al., 2019; Dash et al. , 2018; Banadkooki et al. , 2019; Mehr et al., 2019; Damavandi and Shah, 2019; Lee et al., 2018; Beheshti et al., 2016; Nourani et al., 2019; Abbot and Marohasy, 2017; Kumar et al., 2019; Haidar and Verma, 2018; Gao et al., 2019; Mishra and Kushwaha, 2019; Zhang et al., 2020, 2018; Yu et al., 2017; Zainudin et al., 2016; Diez-Sierra and del Jesus, 2020; Singh and Kumar, 2019; Oswal, 2019; Aguasca-Colomo et al., 2019; Peng et al., 2019; Pham et al., 2019; Valencia-Payan and Corrales, 2018; Balamurugan and Manojkumar, 2021; Chen et al., 2016; Du et al., 2017; Manandhar et al., 2019; Kashiwao et al., 2017; Huang et al., 2017; Pan et al., 2019; Du et al., 2018; Zhan et al., 2019; Chattopadhyay et al., 2020; Patel et al., 2018; Zhuang and Ding, 2016; Boonyuen et al., 2018, 2019), with the remaining 12 studies producing a series 2D images as output (Shi et al., 2015; Singh et al., 2017; Jing et al., 2019; Sato et al., 2018; Ayzelet al., 2019; Chen et al., 2020; Tran and Song, 2019a; Shi et al., 2017; Castro et al., 2020; Wang et al., 2017; Tran and Song, 2019b; Shi et al., 2017). Of the studies with 2D output, all except one (Castro et al., 2020) involve short-term prediction intervals of 10 minutes or less.

A connection between forecasting time frame and the discrete/continuous nature of the output can be observed. In general, studies involving longer-term predictions tend to make use of regression which produces continuous output, whereas on short-term time frames, studies tend towards using classification that gives discrete output. Specifically, 27 out of the 30 papers that focus on long-term prediction carry out regression (Du et al., 2019; Dash et al. , 2018; Cristian, 2018; Lakshmaiah et al., 2016; Ramsundram et al., 2016; Sulaiman and Wahab, 2018; Amiri et al., 2016; Abbot and Marohasy, 2016; Sardeshpande and Thool, 2019; Shenify et al., 2016; Banadkooki et al. , 2019; Mehr et al., 2019; Damavandi and Shah, 2019; Lee et al., 2018; Beheshti et al., 2016; Nourani et al., 2019; Abbot and Marohasy, 2017; Kumar et al., 2019; Haidar and Verma, 2018; Duong et al., 2018; Xu et al., 2020; Aswin et al., 2018; Canchala et al., 2020; Bojang et al., 2020; Chhetri et al., 2020; Mehdizadeh et al., 2018; Weesakul et al., 2018), and 23 of the 36 papers that focus on short-term prediction carry out classification (Zainudin et al., 2016; Diez-Sierra and del Jesus, 2020; Singh and Kumar, 2019; Oswal, 2019; Balamurugan and Manojkumar, 2021; Chen et al., 2016; Du et al., 2017; Manandhar et al., 2019; Kashiwao et al., 2017; Huang et al., 2017; Shi et al., 2015; Singh et al., 2017; Jing et al., 2019; Sato et al., 2018; Ayzelet al., 2019; Chen et al., 2020; Shi et al., 2017; Zhan et al., 2019; Chattopadhyay et al., 2020; Patel et al., 2018; Zhuang and Ding, 2016; Boonyuen et al., 2018, 2019). This relation may be explained by the fact that longer-term studies usually aim at predicting averages over several days (up to a month), while short-term studies predict instantaneous conditions. Multi-day averaged data assumes a continuous range of values, while in instantaneous rainfall data sets most values are null. It follows that classification into rain/no rain is useful for short term, but not for long-term prediction.

4.5 Input Features

In order to make future predictions, studies make use of data from one or more time steps (called "lags" or "time lags") as input features to predict one or more future lags. For example, to predict rainfall at lag T, two previous time lags $(T-1)$ and $(T-2)$ may be used.

The actual input features in each lag vary across studies. In general, the input features used in the studies in this review were found to be of two types: 1D input features in which each time lag in the data set represents one or a set of geophysical parameters that have been collected at static known locations, i.e., meteorological stations; and 2D input features in which each time lag in the data set is a 2D spatial map of values representing rainfall in the geographical area under review, usually collected by satellite or radar.

1D input features used include geophysical parameters such as temperature, humidity, wind speed and air pressure (Ramsundram et al., 2016; Amiri et al., 2016; Banadkooki et al. , 2019; Damavandi and Shah, 2019; Aguasca-Colomo et al., 2019; Oswal, 2019; Manandhar et al., 2019; Kashiwao et al., 2017; Lu et al., 2018; Xu et al., 2020). In a smaller number of cases, climatic indices such as the Pacific Decadal Oscillation may also be used (Du et al., 2019; Abbot and Marohasy, 2016; Lee et al., 2018; Gao et al., 2019; Valencia-Payan and Corrales, 2018). Studies that use 1D input features tend to use a relatively small number of overall input features, ranging from 2–12 features used for prediction.

With 2D input features, one or more images are taken as input features, depending on the number of time lags used as input, e.g., two time lags used as input implies that two images are used as input. The number of time lags used as input is henceforth referred to as the "sequence length."

There is no rule of thumb for how many time lags should be used as input, and this is mostly selected arbitrarily, and in fewer cases via trial and error. The vast majority of the studies under review select a fixed sequence length. The sequence length can be viewed as a hyper-parameter that affects the prediction outcome, but the optimization of this hyper-parameter is not investigated in the studies under review. The studies under review were found to be more focused on the machine learning component, mostly at devising new deep learning architectures, than selecting and tuning other aspects of their systems.

The most common sequence lengths used are 5 frames (Singh et al., 2017; Jing et al., 2019; Sato et al., 2018; Wang et al., 2017) and 10 frames (Singh et al., 2017; Jing et al., 2019; Sato et al., 2018; Wang et al., 2017). Other sequence lengths are also used, such as 2 (Singh et al., 2017), 4 (Sato et al., 2018), 7 (Wang et al., 2017) and 20 (Jing et al., 2019).

Studies that use 2D input features tend to use a relatively large number of input features. This can be attributed to the fact that the feature vectors produced are associated with one or more 2D images, resulting in vectors of size (Image width $\times$ Image height $\times$ Sequence length). Overall, the number of features can grow as high as several thousands.

Typically, 1D or 2D inputs are used to predict 1D or 2D outputs, respectively. As noted in the previous section, longer-term predictions tend to make 1D predictions, so it follows these studies also tend to use 1D data (Du et al., 2019; Dash et al. , 2018; Cristian, 2018; Lakshmaiah et al., 2016; Ramsundram et al., 2016; Sulaiman and Wahab, 2018; Amiri et al., 2016; Abbot and Marohasy, 2016; Sardeshpande and Thool, 2019; Banadkooki et al. , 2019; Mehr et al., 2019; Damavandi and Shah, 2019; Lee et al., 2018; Beheshti et al., 2016; Nourani et al., 2019; Abbot and Marohasy, 2017; Kumar et al., 2019; Duong et al., 2018; Xu et al., 2020; Canchala et al., 2020; Bojang et al., 2020; Chhetri et al., 2020; Mehdizadeh et al., 2018; Mishra and Kushwaha, 2019; Aguasca-Colomo et al., 2019), while those that make shorter-term predictions tend towards the use of 2D data (Shi et al., 2015; Singh et al., 2017; Jing et al., 2019; Sato et al., 2018; Ayzelet al., 2019; Chen et al., 2020; Tran and Song, 2019a; Shi et al., 2017; Castro et al., 2020; Wang et al., 2017; Tran and Song, 2019b; Shi et al., 2017; Boonyuen et al., 2019)

4.6 Input Data Pre-processing

Before ML tools are applied to make predictions on the available data, the input data is usually pre-processed to reformat the data into a form that will make training of, and prediction by, the ML tool(s) easier and faster. The pre-processing techniques usually applied in geophysical parameter forecasting can be broken down into three broad categories, namely data imputation; feature selection/reduction; and data preparation for classification. The following subsections describe these categories, as well as their application in the papers in this review.

4.6.1 Data Imputation

Data sets are regularly found to have missing data entries, which is caused by a range of factors such as data corruption, data sensor malfunction, etc. This is a serious issue faced by researchers in data mining or analysis and needs to be addressed as part of pre-processing before feature selection/preparation and training.

The techniques used to infer and substitute missing data are collectively referred to as data imputation techniques. Data imputation is challenging and is an on-going research area. In the papers in this review, it was found that very little focus was placed on this problem, with most of the studies making use of simple statistical techniques such as averaging to interpolate missing data entries (Sulaiman and Wahab, 2018; Haidar and Verma, 2018; Canchala et al., 2020; Bojang et al., 2020; Zainudin et al., 2016; Oswal, 2019). While not used in the papers in this review, more advanced data imputation techniques exist beyond the use of simple statistics, such as the use of ML to impute the data. The interested reader may refer to (Tang and Ishwaran, 2017; Shah et al., 2014; Pantanowitz and Marwala, 2009).

4.6.2 Feature Selection/Reduction

Feature selection/reduction aims to determine and use salient features in the data, and disregard irrelevant features in the data. This helps to reduce training time, decrease the model complexity and increase its performance. In the papers in this review, it is observed that feature selection is carried out either automatically or manually.

For automatic feature selection, various algorithms are used to determine the most salient features in the data. The most common method used in the papers in this review involved the use of deep learning techniques such as ANNs and convolutional neural networks (CNNs), to select/reduce features automatically, most especially when high-dimensional data such as radar and satellite images was used (Shi et al., 2015; Singh et al., 2017; Jing et al., 2019; Sato et al., 2018; Chen et al., 2020; Shi et al., 2017; Castro et al., 2020; Wang et al., 2017; Tran and Song, 2019b; Shi et al., 2017; Patel et al., 2018; Zhuang and Ding, 2016; Boonyuen et al., 2018, 2019; Haidar and Verma, 2018; Weesakul et al., 2018). The use of deep learning techniques was found to be much more common with short-term data sets which are generally much larger, therefore making it possible to achieve convergence on deep networks. Another category of ML tools used for automatic feature selection includes ensemble methods like random forests (RFs) which automatically order features in terms of importance, as used in Damavandi and Shah (2019), Bojang et al. (2020); Aguasca-Colomo et al. (2019), Valencia-Payan and Corrales (2018), Yu et al. (2017), Zainudin et al. (2016), Diez-Sierra and del Jesus (2020), Singh and Kumar (2019), and Balamurugan and Manojkumar (2021). Finally, principle components analysis (PCA) has also been used to reduce features in Du et al. (2019), Peng et al. (2019), Zhang et al. (2018), Diez-Sierra and del Jesus (2020), Du et al. (2017), and Pan et al. (2019).

As regards manual feature selection, researchers may either use prior experience and trial and error to manually select relevant features such as in Dash et al. (2018), Cristian (2018), Dash et al. (2018), Cristian (2018), Lakshmaiah et al. (2016), Sulaiman and Wahab (2018), Sardeshpande and Thool (2019), Shenify et al. (2016), Banadkooki et al. (2019), Nourani et al. (2019), Haidar and Verma (2018), Duong et al. (2018), and Canchala et al. (2020) or use correlation analysis methods such as auto correlation to indirectly inform the manual feature selection process as in Du et al. (2019), Ramsundram et al. (2016), Abbot and Marohasy (2016), Mehr et al. (2019), Damavandi and Shah (2019), Lee et al. (2018), Abbot and Marohasy (2017), Kumar et al. (2019), and Gao et al. (2019). Where images are used, image cropping and resizing is applied to, respectively, dispose of irrelevant/static image segments and reduce the number of features (Shi et al., 2015; Singh et al., 2017; Jing et al., 2019; Sato et al., 2018; Chen et al., 2020; Shi et al., 2017; Castro et al., 2020; Wang et al., 2017; Tran and Song, 2019b; Shi et al., 2017).

Manual feature selection is much more common with long-term data sets, with very few long-term prediction studies in this review making use of automatic feature selection methods. This is partly attributed to the relatively smaller amount of data

available in these sets, as mentioned before, which makes it challenging, or even rules out, the application of, e.g., deep learning methods for automatic feature selection.

The rotation of the earth around the sun can cause data to exhibit a seasonal behavior on an annual basis, i.e., they exhibit annual periodicity (Delleur and Kavvas, 1978; Barnettet al., 2012). This is most prominent in long-term data sets and less prominent in shorter-term data sets. Addressing seasonality in long-term data sets is critical when traditional time series models are used, since these models assume stationarity (Delleur and Kavvas, 1978; Nielsen, 2020), while seasonality and trends in general makes time series non-stationary. Converting data from a non-stationary to a stationary state involves is a process of generating a time series with statistical properties that do not change over time. For further information about seasonal and non-stationary data sets and the conversion of non-stationary to stationary time series, the interested reader is referred to (Nielsen, 2020). Another way to deal with seasonality is the inclusion of features that exhibit seasonal behavior, such as the usage of the same month previous year.

Figure 4.5 shows the methodologies used in the long-term prediction studies in this review. 11 of the 30 long-term papers (37%) did not address seasonality in the data (Du et al., 2019; Dash et al. , 2018; Ramsundram et al., 2016; Sardeshpande and Thool, 2019; Banadkooki et al. , 2019; Damavandi and Shah, 2019; Lee et al., 2018; Mehdizadeh et al., 2018; Mishra and Kushwaha, 2019; Aguasca-Colomo et al., 2019; Gao et al., 2019),while the remaining 19 papers used some means of addressing seasonality in the data (Bojang et al., 2020; Amiri et al., 2016; Shenify et al., 2016; Xu et al., 2020; Mehr et al., 2019; Beheshti et al., 2016; Canchala et al., 2020; Cristian, 2018; Lakshmaiah et al., 2016; Sulaiman and Wahab, 2018; Abbot and Marohasy, 2016; Nourani et al., 2019; Abbot and Marohasy, 2017; Kumar et al., 2019; Haidar and Verma, 2018; Duong et al., 2018; Aswin et al., 2018; Chhetri et al., 2020; Weesakul et al., 2018).

In the papers that addressed seasonality, four unique approaches were identified, and some were combined with others. The first approach involves including features from lag $T-12$ (same month previous year) in the feature set used to predict rainfall at month T (Mehr et al., 2019; Beheshti et al., 2016; Canchala et al., 2020; Cristian, 2018; Lakshmaiah et al., 2016; Sulaiman and Wahab, 2018; Abbot and Marohasy, 2016; Nourani et al., 2019; Abbot and Marohasy, 2017; Kumar et al., 2019; Haidar and Verma, 2018; Duong et al., 2018; Aswin et al., 2018; Chhetri et al., 2020; Weesakul et al., 2018). A less common approach is to use the index of the current month in the year (1=January, ..., 12=December) as an input feature (Haidar and Verma, 2018; Beheshti et al., 2016).

Alternative approaches include performing time series decomposition, either using singular spectrum analysis as in Bojang et al. (2020) or wavelet transformation as in Amiri et al. (2016), Xu et al. (2020), and Shenify et al. (2016). One paper (Beheshti et al., 2016) combined time series decomposition using singular spectrum analysis with the inclusion of features from lag $T-12$ in the feature set. This has been included in the segment labelled “Combination" in Fig. 4.5.

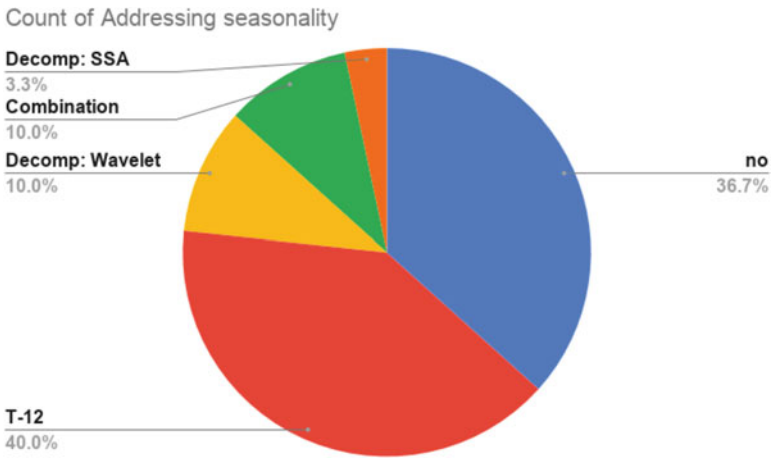

Fig. 4.5 Methods used to account for seasonality in studies with long-term data, by percentage

The final approach used to address seasonality takes the form of data de-seasonalization by subtracting the monthly averages from the data as in Chhetri et al. (2020) and Mehr et al. (2019). All of the papers in this review that used this approach combined this subtraction with the first approach, i.e., including features from lag $T - 12$ in the feature set. These two papers have also been included in the segment labelled “Combination" in Fig. 4.5.

4.6.3 Data Preparation for Classification

When attempting to carry out classification into discrete classes, it is either necessary to use a data set in which the desired output variable is discrete or to convert a desired continuous-valued output variable into discrete classes. This involves setting the desired number of classes, which is usually done manually and arbitrarily, followed by determining the range of values represented by each class, i.e., determining the thresholds that divide the continuous scale into the desired classes. Finally, where the number of instances across classes is imbalanced, it is necessary to balance them.

In the papers in this survey that carried out classification, most made use of data that was continuous, yet very few provide details on the process used to convert from a continuous to a discrete scale. Select studies in this survey that provide information about their data preparation process are described below.

In converting from continuous to discrete data, after manually specifying the number of classes (which has been explained in Sect. 4.4), studies automate the selection of the class thresholds using clustering tools, specifically k-means and k-medoids (Chattopadhyay et al., 2020; Mishra and Kushwaha, 2019; Aguasca-

Colomo et al., 2019). Another approach taken is to manually determine suitable thresholds, by performing a series of experiments to compare various threshold values (Shi et al., 2017). To address any resulting class imbalances, researchers may perform random down-sampling to obtain an equal sample distribution across classes as in Aguasca-Colomo et al. (2019), Oswal (2019), and Manandhar et al. (2019).

4.7 Machine Learning Techniques Used

The studies in this survey made use of a wide range of ML techniques which can be subdivided into two main groups: "classical" techniques such as multivariate linear regression (MLR), KNN ANNs, SVMs, and RF; and modern deep learning methods such as CNNs and Long-Short-Term-Memory (LSTM). It was observed that classical ML models tended to work with 1D data from meteorological stations, such as in Mallika and Nirmala (2016), Du et al. (2019), Dash et al. (2018), Amiri et al. (2016), Pham et al. (2019), Yu et al. (2017), and Ramsundram et al. (2016) for short-term data and Du et al. (2019), Cristian (2018), Lakshmaiah et al. (2016), Ramsundram et al. (2016), Abbot and Marohasy (2016), Shenify et al. (2016), Mehr et al. (2019), Damavandi and Shah (2019), Mehdizadeh et al. (2018), Gao et al. (2019), and Aguasca-Colomo et al. (2019) for long-term data.

Some papers use hybrid models that combine two or more approaches. A popular hybrid approach is to combine ML with optimization tools such as genetics and particle swarm optimization to optimize hyper-parameters (Mehdizadeh et al., 2018; Beheshti et al., 2016; Mehr et al., 2019; Du et al., 2018, 2017). Multiple ML techniques are combined in Chhetri et al. (2020), Singh and Kumar (2019), and Peng et al. (2019), and ML is used with ARIMA in Pham et al. (2019).

Deep learning models usually requires huge data sets to avoid overfitting on the data, which explains their popularity among short-term data sets, especially those using 2D data (Shi et al., 2015; Singh et al., 2017; Jing et al., 2019; Sato et al., 2018; Ayzelet al., 2019; Chen et al., 2020; Tran and Song, 2019a; Shi et al., 2017; Castro et al., 2020; Wang et al., 2017; Tran and Song, 2019b; Shi et al., 2017; Pan et al., 2019; Du et al., 2018; Zhan et al., 2019; Chattopadhyay et al., 2020; Patel et al., 2018; Zhuang and Ding, 2016; Boonyuen et al., 2018, 2019). 2D data, in particular, has a huge feature space, which requires authors to implement automated feature reduction models like CNNs (Zhan et al., 2019; Chattopadhyay et al., 2020; Patel et al., 2018; Zhuang and Ding, 2016; Boonyuen et al., 2018).

In order to accommodate the time dimension in the data, many researchers try to adapt time series models such as LSTMs for 1D data in Kumar et al. (2019), Duong et al. (2018), Xu et al. (2020), Aswin et al. (2018), Canchala et al. (2020), Zhang et al. (2020), and Patel et al. (2018). For 2D data, models combining CNNs with LSTMs (designated as ConvLSTMs models) were first used in inShi et al. (2015) in 2015, and subsequently several variations have been implemented (Singh et al.,

2017; Jing et al., 2019; Sato et al., 2018; Ayzelet al., 2019; Chen et al., 2020; Tran and Song, 2019a; Shi et al., 2017; Castro et al., 2020; Wang et al., 2017; Tran and Song, 2019b; Shi et al., 2017).

4.8 Reporting of Results and Accuracy Measures

Several different metrics are used in the literature to measure the performance of the ML models according to the type of the problem. In classification problems, authors tend use metrics such as precision, recall, and accuracy (Gao et al., 2019; Mishra and Kushwaha, 2019; Aguasca-Colomo et al., 2019; Zainudin et al., 2016; Diez-Sierra and del Jesus, 2020; Singh and Kumar, 2019; Oswal, 2019; Chen et al., 2016; Manandhar et al., 2019; Kashiwao et al., 2017; Huang et al., 2017; Singh et al., 2017; Zhan et al., 2019; Chattopadhyay et al., 2020; Patel et al., 2018; Zhuang and Ding, 2016; Boonyuen et al., 2018). If the data is not balanced then f1-score is used rather than the accuracy, since accuracy does not take the imbalance between the classes into account (Zainudin et al., 2016; Diez-Sierra and del Jesus, 2020; Singh and Kumar, 2019; Oswal, 2019; Singh et al., 2017; Patel et al., 2018). For sequence classification prediction, other metrics are used such as the critical success (CSI) (Sato et al., 2018; Ayzelet al., 2019; Chen et al., 2020; Tran and Song, 2019a; Shi et al., 2017; Tran and Song, 2019b; Shi et al., 2017). For continuous outputs, then the mean absolute error, and the root mean squared error are the most commonly used metrics in the literature (Peng et al., 2019; Pham et al., 2019; Valencia-Payan and Corrales, 2018; Zhang et al., 2020, 2018; Yu et al., 2017; Du et al., 2019; Dash et al. , 2018; Cristian, 2018; Lakshmaiah et al., 2016; Ramsundram et al., 2016; Sulaiman and Wahab, 2018; Amiri et al., 2016; Abbot and Marohasy, 2016; Sardeshpande and Thool, 2019; Shenify et al., 2016; Banadkooki et al. , 2019; Mehr et al., 2019; Damavandi and Shah, 2019; Lee et al., 2018; Beheshti et al., 2016; Nourani et al., 2019; Abbot and Marohasy, 2017; Kumar et al., 2019; Haidar and Verma, 2018; Duong et al., 2018; Xu et al., 2020; Aswin et al., 2018; Canchala et al., 2020; Bojang et al., 2020; Chhetri et al., 2020; Mehdizadeh et al., 2018)

A direct comparison of these results across different papers is a nearly impossible task, since each paper uses its own models, pre-processing, metrics, data sets and parameters. However, individual authors frequently compare multiple algorithms, and there are a few ML algorithms that stand out as being most frequently mentioned as better performers. ANNs and deep learning are most frequently mentioned as best performing models, for both long-term prediction (Du et al., 2019; Lakshmaiah et al., 2016; Amiri et al., 2016; Abbot and Marohasy, 2016; Sardeshpande and Thool, 2019; Banadkooki et al. , 2019; Abbot and Marohasy, 2017; Weesakul et al., 2018; Aswin et al., 2018; Kumar et al., 2019; Haidar and Verma, 2018; Duong et al., 2018; Canchala et al., 2020; Chhetri et al., 2020; Mehdizadeh et al., 2018) and especially for short-term prediction (Shi et al., 2015; Singh et al., 2017; Jing et al., 2019; Sato et al., 2018; Ayzelet al., 2019; Chen et al., 2020; Tran and Song, 2019a; Shi et al., 2017; Castro et al., 2020; Wang et al., 2017; Tran and Song, 2019b; Shi et al., 2017;

Pan et al., 2019; Du et al., 2018; Zhan et al., 2019; Chattopadhyay et al., 2020; Patel et al., 2018; Zhuang and Ding, 2016; Boonyuen et al., 2018, 2019; Zhang et al., 2020, 2018; Diez-Sierra and del Jesus, 2020; Kashiwao et al., 2017).

Other algorithms mentioned as best performers are SVMs in 6 studies (Yu et al., 2017; Manandhar et al., 2019; Chen et al., 2016; Du et al., 2017; Shenify et al., 2016; Mehr et al., 2019) ensemble in Valencia-Payan and Corrales (2018), Zainudin et al. (2016), Ramsundram et al. (2016), Nourani et al. (2019), Xu et al. (2020), and Aguasca-Colomo et al. (2019), logistic regression in three studies (Oswal, 2019; Balamurugan and Manojkumar, 2021; Gao et al., 2019) and KNNs in two studies (Huang et al., 2017; Cristian, 2018).

4.9 Discussion

The above sections clearly demonstrate that there is a robust, growing literature on rainfall prediction, which covers an extremely wide variety of time scales, features used, pre-processing techniques, and ML algorithms used. From a high-level perspective, the field can be divided into short versus long time scales (time intervals of a one day or less, versus intervals of a month or more), which tend to have divergent characteristics.

Short-term studies typically rely on huge data sets, and require deep learning applied to large feature sets to find hidden patterns in those data sets. On the other hand, long-term studies rely more on pre-processing methods such as feature selection, data imputation, and data balancing in order to make effective predictions. ANNs and deep learning seem are becoming increasingly prevalent in long-term studies as well as short term: since 2018, 7 of 23 papers on long-term prediction utilized deep learning tools.

There are reasons to regard the trend towards more complicated models with skepticism. Some recent studies have shown that much simpler models such as KNNs can sometimes outperform advanced ML techniques like RNNs, (Lin, 2019; Hussein et al., 2020; Ludewig and Jannach, 2018; Cristian, 2018; Hussein et al., 2021). Similar findings have been reported for other ML applications, such as the top n recommendation problem (Dacrema et al., 2019).

These results underscore the importance of providing simple but statistically well-motivated baselines to verify whether ML truly is effective in improving predictive accuracy. However, many papers do not provide simple baselines, but rather compare several variations or architectures of more advanced ML methods such as SVR or MLP (Mohamadi et al., 2020; Bojang et al., 2020; Mehdizadeh et al., 2018; Canchala et al., 2020; Peng et al., 2019; Chen et al., 2016; Du et al., 2017; Manandhar et al., 2019; Du et al., 2018; Zhan et al., 2019; Patel et al., 2018; Boonyuen et al., 2018, 2019). Of the total reviewed papers, almost half (48.2%) of papers did not supply simple baselines. Of those papers that did supply baselines, a variety of methods was used. For short-term image data, the previous image is frequently used as an untrained predictor for the next image (Tran and Song, 2019a;

Shi et al., 2017; Tran and Song, 2019b; Shi et al., 2017). For monthly data, some papers use MLR based on multiple previous lags (Chhetri et al., 2020; Du et al., 2019; Cristian, 2018; Lakshmaiah et al., 2016); while same month averages, though statistically well-motivated, are used much less frequently (Xu et al., 2020).

Besides the issue of baselining, the use of error bars is essential for comparison purposes, as it highlights whether the improvement obtained by the models are significant. Unfortunately most of the literature in ML does not provide error bars around the measured metrics. In the case of our reviewed literature shows that 88% of the papers did not give error bars.

A final issue of concern is data leakage. Data leakage refers to allowing data from the testing set to influence the training set. Data leakage occurs during the pre-processing of the data, and can take various forms as follows:

- Random shuffling, which involves choosing sequences from a common data pool for both training and testing:
- Imputation, which involves filling missing records using statistical methods on the entire data set (including both training and testing)
- De-seasonalization which utilizes the monthly averages from the entire data set.
- Using current lags, e.g., using temperature at a time T to predict rainfall at the same time T. (Depending on the application, this may or may not constitute data leakage)
- Combination: Which uses two of the above-mentioned techniques.

Figure 4.6, shows the reviewed papers in terms of data leakage. The top chart focuses on long-term data, where the bottom focuses on short-term data. We mentioned previously that long-term data often undergoes more pre-processing than short-term data. This reflects on the graph, as leakage-producing methods are more than twice as common for long term as for short term. Random shuffling was performed in Du et al. (2019); Lakshmaiah et al. (2016), Lee et al. (2018), Beheshti et al. (2016), Duong et al. (2018), Gao et al. (2019), Mishra and Kushwaha (2019), and Aguasca-Colomo et al. (2019) for long-term data, and in Valencia-Payan and Corrales (2018), Du et al. (2017), Manandhar et al. (2019), Pan et al. (2019), Du et al. (2018), and Chattopadhyay et al. (2020) for short-term data. Data imputation was performed in Sulaiman and Wahab (2018), Haidar and Verma (2018), Canchala et al. (2020), and Bojang et al. (2020) for long-term data, and in Oswal (2019) for short-term data. Faulty de-seasonalization was carried out in Mehr et al. (2019) for long-term data. Using the current lags was seen only implemented in Ramsundram et al. (2016). Multiple leakage issues (denoted as "combination" in the figure) were observed in Chhetri et al. (2020) and Zainudin et al. (2016).

4.10 Conclusions

In the area of rainfall prediction, 66 relevant papers are reviewed, by examining the data source, output objective, input feature, pre-processing methods, models

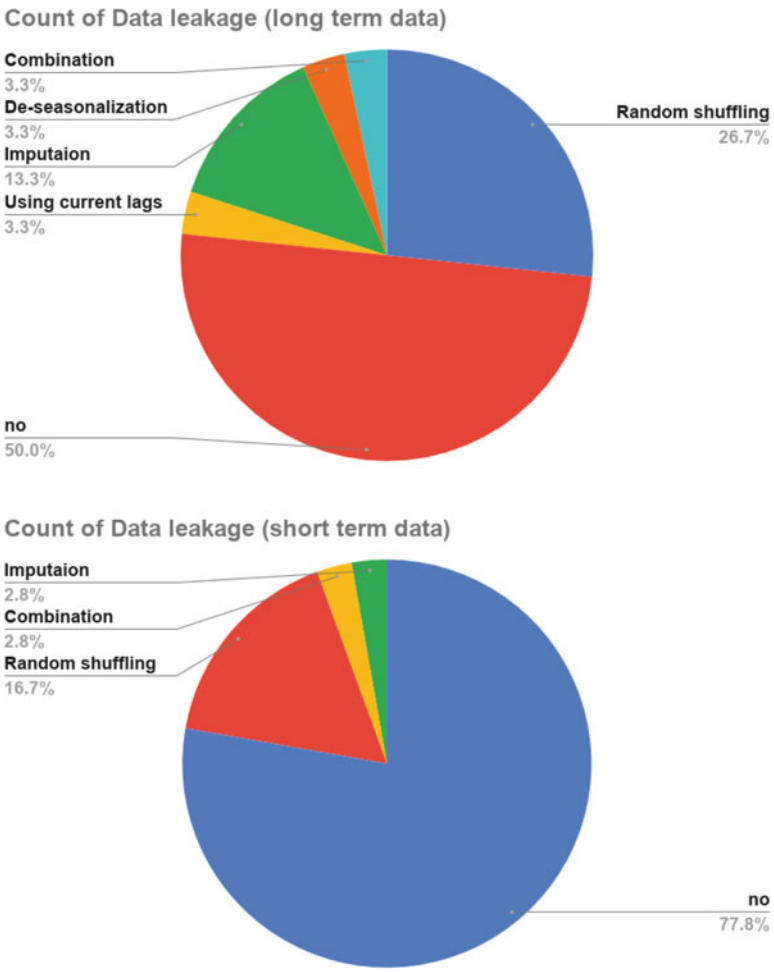

Fig. 4.6 Percentage of papers which introduced data leakage during pre-processing, for long-term data (*top*) and short-term data (*bottom*)

used, and finally the results. Different pre-processing like random shuffling used in the literature suggests that in some cases model performance is inaccurately represented. The aim of the survey is to make researches aware of the different pitfalls that can leads to unreal models performance, which does not only apply for rainfall, but for other time series data.

Acknowledgments E.A.H. acknowledges financial support from the South African National Research Foundation (NRF CSUR Grant Number 121291 for the HIPPO project) and from the Telkom-Openserve-Aria Technologies Center of Excellence at the Department of Computer Science of the University of the Western Cape.

Appendix 1: List of Abbreviations

ML	Machine learning
AD	Author defined
ANNs	Artificial neural networks
CNNs	Convolution neural networks
LSTMs	Long short-term memory
ConvLSTMs	Convolutions layers with Long short-term memory
RF	Random forest
RF	SVMs Support vector machines
DT	Decision tress
XGB	Extreme gradient boosting
LogReg	Logistic regression
MLR	Multi linear regression
KNNs	K-nearest neighbour
RMSE	Root mean square error
MAE	mean absolute error
CA	Classification accuracy
pre	precision
f1	f1-score
PACF	Partial autocorrelation function
ACF	Autocorrelation function
PCA	principle component analysis
NOAA	National Oceanic and Atmospheric Administration

Appendix 2: Summary Tables for References

This appendix contains four tables which summarize the findings for the reviewed papers for long-term data Tables 4.1 and 4.2, and short-term data Tables 4.3 and 4.4. Tables 4.1 and 4.3 contain information regarding the source, period, region, input, output; while Tables 4.2 and4.4 include information about the pre-processing tools, data leakage, and the ML used.

Table 4.1 Data sources, spatio-temporal coverage, inputs and outputs, and references for long-term predictive studies

No.	Source	Period	Region	Input	Output	Ref
1	China meteorological administration (CMA)	1916–2015	China	6 climatic indices	Seasonal regression	Du et al. (2019)
2	Indian institute of tropical meteorology (IITM)	1817–2016	India	8 past lags	Seasonal regression	Dash et al. (2018)
3	Romanian rainfall	1991–2015	Romania	12 past lags	Monthly regression	Cristian (2018)
4	Rainfall from the India water portal	1901–2002	India	11 climatic parameters	Monthly regression	Lakshmaiah et al. (2016)
5	Tuticorin meteorological station	1980–2002	India	Four climatic parameters	Monthly regression	Ramsundram et al. (2016)
6	Malaysian department of irrigation and drainage	1965–2015	Malaysia	10 past lags	Monthly regression	Sulaiman and Wahab (2018)
7	National cartographic center of Iran (NCC)	1996–2010	Iran	Four climatic parameters	Monthly regression	Amiri et al. (2016)
8	Royal Netherlands meteorological institute climate explorer	2004–2014	Australia	Seven climatic indices	Monthly regression	Abbot and Marohasy (2016)
9	Indian water portal	1901–2000	India	Four climatic parameters	Monthly regression	Sardeshpande and Thool (2019)
10	Serbian meteorological stations	1946–2012	Serbia	Past rainfall lags	Monthly regression	Shenify et al. (2016)
11	Iran meteorological department	2000–2010	Iran	Two climatic parameters	Monthly regression	Banadkooki et al. (2019)
12	Iran meteorological department	1990–2014	Iran	Four past lags	Monthly regression	Mehr et al. (2019)
13	CHIRPS, and NCEP-NCAR Reanalysis	1918–2001	Indus basin	5 climatic features	Monthly regression	Damavandi and Shah (2019)
14	World agrometeorological information service (WAMIS) and NOAA	1966–2017	South Korea	11 climatic indices	Monthly regression	Lee et al. (2018)
15	Malaysian department of irrigation and drainage	1950–2010	Malaysia	6 past lags and time stamp	Monthly regression	Beheshti et al. (2016)
16	Turkish stations	2007–2016	Turkey	3 rainfall lags	Monthly regression	Nourani et al. (2019)

(continued)

Table 4.1 (continued)

No.	Source	Period	Region	Input	Output	Ref
17	Australian stations	1885–2014	Australia	10 climatic indices and parameters	Monthly regression	Abbot and Marohasy (2017)
18	Indian meteorological department	1871–2016	India	12 past lags	Monthly regression	Kumar et al. (2019)
19	Bureau of meteorology (BOM), Royal Netherlands meteorological institute climate, more	1908–2012	Australia	43 climatic indices and parameters	Monthly regression	Haidar and Verma (2018)
20	Vietnam's hydrological gauging	1971–2010	Vietnam	12 features	Monthly regression	Duong et al. (2018)
21	Global precipitation climatology center (GPCC)	1901–2013	China	6–9 climatic indices and parameters	Monthly regression	Xu et al. (2020)
22	Precipitation from NCEP	1979–2018	GLOBAL	164 past lags	Monthly regression	Aswin et al. (2018)
23	National center of hydrology and meteorology department (NCHM)	1997–2015	Bhutan	6 climates parameters	Monthly regression	Canchala et al. (2020)
24	Taiwan water resource bureau	1958–2018	Taiwan	3 past lag	Monthly regression	Bojang et al. (2020)
25	Instituto de Hidrología, Meterología y Estudios Ambientales (IDEAM) of Colombia	1983–2016	Colombia	6 past lags	Monthly regression	Chhetri et al. (2020)
26	Islamic Republic of Iran meteorological organization (IRIMO)	1981- 2012	Iran	5 past lags	Monthly regression	Mehdizadeh et al. (2018)
27	Pluak Daeng station in Thailand	1991–2016	Thailand	346 climatic indices and parameters	Monthly regression	Weesakul et al. (2018)
28	National climate center of China meteorological administration (NCC-CMA)	1952–2012	China	84 climatic indices	Yearly Classification	Gao et al. (2019)
29	The department of agricultural meteorology Indira	2011–2013	India	Five climatic parameters	Monthly classification	Mishra and Kushwaha (2019)
30	Meteorological stations of the island of Tenerife and NOAA databases	1976–2016	Tenerife Island	12 climatic indices and parameters	Monthly classification	Aguasca-Colomo et al. (2019)

Table 4.2 Pre-processing, data leakage characteristics, machine learning algorithms used, and reference numbers for long-term predictive studies

No.	Pre-processing	Data leakage	ML used	Ref
1	Normalization, random shuffling, feature correlation	Random shuffling	PCA-ANN, PCA-MLR	Du et al. (2019)
2	Normalization	No	KNNs, ANNs, ELM	Dash et al. (2018)
3	Windowing	No	KNNs, ARIMA, ANNs	Cristian (2018)
4	Windowing, random shuffling	Random shuffling	ANN, ARMA, LR	Lakshmaiah et al. (2016)
5	Data imputation, noise removal, correlation analysis	Using current lags	DT, ANNs	Ramsundram et al. (2016)
6	Normalization, and data imputation	Imputation	ANNs, ARIMA	Sulaiman and Wahab (2018)
7	Normalization, decomposition	no	WTANN, ANNs	Amiri et al. (2016)
8	Features correlation	No	ANNs, POAMA	Abbot and Marohasy (2016)
9	Normalization	No	Different ANNs	Sardeshpande and Thool (2019)
10	N/A	No	ANN, WT-SVM, GP	Shenify et al. (2016)
11	Normalization, optimization	No	AD-MLP, AD-SVM, DT	Banadkooki et al. (2019)
12	Correlation analysis (PACF), square root transformation, standardization, de-seasonalization	De-seasonalization	SVR, AD-SVR, more	Mehr et al. (2019)
13	Feature correlation, random shuffling	No	MLP, SVR, MLR, RF, KNNs	Damavandi and Shah (2019)
14	Feature correlation, random shuffling	Random shuffling	ANNs	Lee et al. (2018)
15	Decomposition	Random shuffling	AD-MLP	Beheshti et al. (2016)
16	Normalization	No	Ensemble method, SVM, ANNS, more	Nourani et al. (2019)
17	Feature selection	No	ANNs, POAMA	Abbot and Marohasy (2017)
18	Feature correlation, windowing	N/A	LSTM, RNN	Kumar et al. (2019)
19	Data imputation, normalization	Imputation	1D-CNN, MLP, baseline (ACCESS-S1)	Haidar and Verma (2018)

(continued)

Table 4.2 (continued)

No.	Pre-processing	Data leakage	ML used	Ref
20	Random shuffling	Random shuffling	MLP, LSTM, SNN	Duong et al. (2018)
21	Normalization, wavelet	No	MLR, MLP, LSTM, SVMs, ConvLSTMs, ensemble methods	Xu et al. (2020)
22	Grey-scale, windowing	No	LSTM, ConvNet	Aswin et al. (2018)
23	Normalization, data imputation	Imputation	MLR , AD-LSTM, LSTM, MLP	Canchala et al. (2020)
24	Decomposition	Imputation	UD-RF, RF, UD-SVR, SVR	Bojang et al. (2020)
25	Imputation, de-seasonlization	Imputation, de-seasonalization	3 AD-ANNs models	Chhetri et al. (2020)
26	Normalization	No	ANNs, AD-ANNs, AD-gene expression programming	Mehdizadeh et al. (2018)
27	N/A	No	DNNs	Weesakul et al. (2018)
28	Feature correlation, feature reduction	Random shuffling	MLogR	Gao et al. (2019)
29	Clustering	Random shuffling	GPR, DT, NB	Mishra and Kushwaha (2019)
30	Random down-sampling, feature correlation	Random shuffling	XGB, RF, more	Aguasca-Colomo et al. (2019)

Table 4.3 Data sources, spatio-temporal coverage, inputs and outputs, and references for short-term predictive studies

No.	Source	Period	Region	Input	Output	Ref
1	Indian statistical institute	1989–1995	Multiple regions	10 climatic parameters	Daily regression	Peng et al. (2019)
2	Vietnamese stations	1978–2016	Vietnam	Previous lags	Daily regression	Pham et al. (2019)
3	Meteoblue data , MODIS, and more	2012–2014	Colombia	12 climatic indices and parameters	Hourly regression	Valencia-Payan and Corrales (2018)
4	Central meteorological observatory of Shanghai	2015–2017	China	24 climatic parameters	Hourly regression	Zhang et al. (2020)
5	China meteorological administration	2015–2017	China	13 climatic parameters	Hourly regression	Zhang et al. (2018)
6	Taiwan and the national severe storms laboratory and NOAA	2012–2015	Taiwan	3–4 parameters	Hourly regression	Yu et al. (2017)
7	Meteorological drainage and the irrigation departments in Malaysia	2010–2014.	Malaysia	4 parameters	Daily classification	Zainudin et al. (2016)
8	The water planning and managing agency for Tenerife Island, and NOAA	1979–2015	Spain	1800 parameters	Daily classification	Diez-Sierra and del Jesus (2020)
9	U.S. government's open data	2010–2017	US	25 parameters	Daily classification	Singh and Kumar (2019)
10	Kaggle and the Australian government	2008–2017	Australia	23 parameters	Daily classification	Oswal (2019)
11	Indian meteorological department	2008–2017	India	8 parameters	Daily classification	Balamurugan and Manojkumar (2021)
12	Satellite imagery data are from FY-2G, and meteorological station located in Shenzhen	2015	China	8 parameters	Hourly classification	Chen et al. (2016)

(continued)

Table 4.3 (continued)

No.	Source	Period	Region	Input	Output	Ref
13	Data from the Nanjing station	N/A	China	6 parameters	Hourly classification	Du et al. (2017)
14	Singapore related weather stations	2012–2015	Singapore	15 climatic parameters	Min classification	Manandhar et al. (2019)
15	Japan meteorological agency	2000–2012	Japan	8 features	Min classification	Kashiwao et al. (2017)
16	NCEP-NCAR and Beijing meteorological station	1990–2012	China	6 climatic indices and parameters	Daily classification	Huang et al. (2017)
17	Radar images collected in Hong Kong	2011–2013	Hong Kong	5 frames	Min classification	Shi et al. (2015)
18	Radar images from USA from 2008–2015	2008–2015	US	10 frames	Min classification	Singh et al. (2017)
19	Radar images from national meteorological information center	2016–2017	China	10 frames	Min classification	Jing et al. (2019)
20	Radar images are retrieved using Yahoo! static map API	2013–2017	Japan	10 frames	Min classification	Sato et al. (2018)
21	Radar images from the German weather service (DWD)	2006–2017	Germany	2 frames	Min both	Ayzelet al. (2019)
22	Weather surveillance radar-1988 doppler radar (WSR-88D)	2015–2018	China	20 frames	Min classification	Chen et al. (2020)
23	CIKM AnalytiCup 2017 competition	N/A	China	5 frames	Min regression	Tran and Song (2019a)
24	CINRAD-SA type Doppler weather radar	2016	China	4 frames	Min classification	Shi et al. (2017)
25	CHIRPS	1918–2019	China	5 frames	Daily regression	Castro et al. (2020)

26	Radar images collected in Hong Kong	2011–2013	China	10 frames	Min regression	Wang et al. (2017)
27	CIKM AnalytiCup 2017 competition	N/A	China	7 frames	Min both	Tran and Song (2019b)
28	Dataset from HKO-7	2009–2015	Hong Kong	5 frames	Min regression	Shi et al. (2017)
29	NCEP, and NOAA	1979–2017	US	A tensor of $8 \times 4 \times 25 \times 25$	Daily regression	Pan et al. (2019)
30	China meteorological data network	N/A	China	7 climatic parameters	Hourly regression	Du et al. (2018)
31	NOAA	1800–2017	US	30 climatic parameters	Hourly both	Zhan et al. (2019)
32	Large ensemble (LENS) community project	1920–2005	US	$3 \times 28 \times 28 \times 3$	Hourly classification	Chattopadhyay et al. (2020)
33	Kaggle	2012–2017	US and India	120 climatic lags	Hourly classification	Patel et al. (2018)
34	Iowa state	1948–2010	USA	9 climatic parameters	Daily classification	Zhuang and Ding (2016)
35	Meteorological department of Thailand and the petroleum authority of Thailand	2017–2017	Thailand	One image	Daily classification	Boonyuen et al. (2018)
36	Meteorological department of Thailand and the petroleum authority of Thailand	2017–2018	Thailand	One and batch of images	Daily classification	Boonyuen et al. (2019)

Table 4.4 Pre-processing, data leakage characteristics, machine learning algorithms used, and reference numbers for short-term predictive studies

No.	Pre-processing	Data leakage	ML used	Ref
1	Normalization, cross validation, feature reduction (PCA)	No	AD-ELM	Peng et al. (2019)
2	Normalization, feature correlation	No	ARIMA-MLP, ARIMA-SVM, ARIMA-HW, ARIMA-NF, more	Pham et al. (2019)
3	Data imputation, data shuffling	Random shuffling	RF, cubist	Valencia-Payan and Corrales (2018)
4	Feature selection, correlation analysis, interpolation, clustering	No	LSTM, MLR, SVMs, ECMFWF	Zhang et al. (2020)
5	Normalization, feature reduction (PCA)	No	DRCF, ARIMA, more	Zhang et al. (2018)
6	N/A	No	RF, SVM	Yu et al. (2017)
7	Normalization, data imputation, shuffling	Data imputation, random shuffling	SVM, RF, DT, NB, ANN	Zainudin et al. (2016)
8	Feature reduction (PCA)	No	ANNs, RF, KNNs, LogR	Diez-Sierra and del Jesus (2020)
9	Feature selection (RF), k-fold cross validation	No	RF,AD[ANNs, Adaboost, SVM, KNN]	Singh and Kumar (2019)
10	Feature selection, feature correlation, data imputation, over, and down-sampling	Imputation	LogR, DT, KNNs, more	Oswal (2019)
11	N/A	No	LogReg, DT, RF, more	Balamurugan and Manojkumar (2021)
12	Radiometric, and geometric correction, and windowing	No	SVM	Chen et al. (2016)
13	Normalization, random shuffling	Random shuffling	AD-SVMs	Du et al. (2017)
14	Down-sampling, feature correlation	Random shuffling	SVM	Manandhar et al. (2019)

15	Outliers removal, normalization	No	MLP, RBFN	Kashiwao et al. (2017)
16	Normalization	No	Knns	Huang et al. (2017)
17	Feature reduction, noise removal, windowing	No	ConvLSTM, FC-LSTM, more	Shi et al. (2015)
18	Resizing, windowing	No	Eulerian persistence, AD-Conv-RNN, ConvLSTM	Singh et al. (2017)
19	Feature reduction, windowing	No	MLC-LSTM, ConvLSTM, more	Jing et al. (2019)
20	Feature reduction, windowing	No	SDPredNet, TrajGRU, more	Sato et al. (2018)
21	Logarithmic transformation	No	Optical flow, Dozhdya.Net	Ayzelet al. (2019)
22	Noise removal, remove corrupted images, windowing, Normalization	No	COTREC, ConvLSTM, AD-ConvLSTM, more	Chen et al. (2020)
23	Normalization, windowing	No	Last frame, TrajGRU, ConvLSTM, AD-TrajGRU, more	Tran and Song (2019a)
24	Windowing, grey-scale transformation	No	Last input, COTREC, AD-CNN	Shi et al. (2017)
25	Windowing, grey-scale, resizing	No	ConvLSTM, AD-ConvLSTMs	Castro et al. (2020)
26	Windowing, grey-scale, resizing	No	ConvLSTM, PredRNN, VPNbaseline	Wang et al. (2017)
27	Windowing, grey-scale, resizing, data augmentation	No	ConvLSTM, ConvGRU, TrajGRU, PredRNN, PredRNN++, last frame	Tran and Song (2019b)
28	Windowing, grey-scale, noise removal, normalization	No	2D CNN, 3D CNN, ConvGRU, TrajGRU, last frame, more	Shi et al. (2017)
29	Normalization, random shuffling	Random shuffling	LR, CNNs, base model (NARR)	Pan et al. (2019)
30	Random shuffling and normalization	Random shuffling	DBN, GA-SVM, more	Du et al. (2018)
31	N/A	No	CNN, LPBoost, more	Zhan et al. (2019)
32	Clustering, down-sampling, random shuffling	Random shuffling	CNN, LogReg	Chattopadhyay et al. (2020)
33	Normalization	No	CNN, LSTM	Patel et al. (2018)
34	Cropping	No	CNN	Zhuang and Ding (2016)
35	Cropping	N/A	CNN	Boonyuen et al. (2018)
36	Cropping	N/A	CNN	Boonyuen et al. (2019)

References

Abbot, J., & Marohasy, J. (2016). Forecasting monthly rainfall in the western Australian wheat-belt up to 18-months in advance using artificial neural networks. In *Australasian Joint Conference on Artificial Intelligence* (pp. 71–87). Berlin: Springer.

Abbot, J., & Marohasy, J. (2017). Application of artificial neural networks to forecasting monthly rainfall one year in advance for locations within the Murray Darling basin, Australia. *International Journal of Sustainable Development and Planning, 12*(8), 1282–1298.

Aguasca-Colomo, R., Castellanos-Nieves, D., & Méndez, M. (2019). Comparative analysis of rainfall prediction models using machine learning in islands with complex orography: Tenerife island. *Applied Sciences, 9*(22), 4931.

Amiri, M. A., Amerian, Y., & Mesgari, M. S. (2016). Spatial and temporal monthly precipitation forecasting using wavelet transform and neural networks, Qara-Qum catchment, Iran. *Arabian Journal of Geosciences, 9*(5), 421.

Aswin, S., Geetha, P., & Vinayakumar, R. (2018). Deep learning models for the prediction of rainfall. In *2018 International Conference on Communication and Signal Processing (ICCSP)* (pp. 0657–0661). Piscataway: IEEE.

Ayzel, G., Heistermann, M., Sorokin, A., Nikitin, O., & Lukyanova, O. (2019). All convolutional neural networks for radar-based precipitation nowcasting. *Procedia Computer Science, 150*, 186–192.

Balamurugan, M. S., & Manojkumar, R. (2021). Study of short term rain forecasting using machine learning based approach. *Wireless Networks, 27*, 5429–5434.

Banadkooki, F. B., Ehteram, M., Ahmed, A. N., Fai, C. M., Afan, H. A., Ridwam, W. M., Sefelnasr, A., & El-Shafie, A. (2019). Precipitation forecasting using multilayer neural network and support vector machine optimization based on flow regime algorithm taking into account uncertainties of soft computing models. *Sustainability, 11*(23), 6681.

Barnett, A. G., Baker, P., & Dobson, A. (2012). Analysing seasonal data. *R Journal, 4*(1), 5–10.

Beheshti, Z., Firouzi, M., Shamsuddin, S. M., Zibarzani, M., & Yusop, Z. (2016). A new rainfall forecasting model using the CAPSO algorithm and an artificial neural network. *Neural Computing and Applications, 27*(8), 2551–2565.

Bojang, P. O., Yang, T.-C., Pham, Q. B., & Yu, P.-S. (2020). Linking singular spectrum analysis and machine learning for monthly rainfall forecasting. *Applied Sciences, 10*(9), 3224.

Boonyuen, K., Kaewprapha, P., & Srivihok, P. (2018). Daily rainfall forecast model from satellite image using convolution neural network. In *2018 IEEE International Conference on Information Technology* (pp. 1–7).

Boonyuen, K., Kaewprapha, P., Weesakul, U., & Srivihok, P. (2019). Convolutional neural network inception-v3: A machine learning approach for leveling short-range rainfall forecast model from satellite image. In *International Conference on Swarm Intelligence* (pp. 105–115). Berlin: Springer.

Canchala, T., Alfonso-Morales, W., Carvajal-Escobar, Y., Cerón, W. L., & Caicedo-Bravo, E. (2020). Monthly rainfall anomalies forecasting for southwestern Colombia using artificial neural networks approaches. *Water, 12*(9), 2628.

Castro, R., Souto, Y. M., Ogasawara, E., Porto, F., & Bezerra, E. (2020). STConvS2S: Spatiotemporal convolutional sequence to sequence network for weather forecasting. *Neurocomputing, 426*, 285–298.

Chattopadhyay, A., Hassanzadeh, P., & Pasha, S. (2020). Predicting clustered weather patterns: A test case for applications of convolutional neural networks to spatio-temporal climate data. Sci. Rep. **10**(1), 1–13.

Chen, L., Cao, Y., Ma, L., & Zhang, J. (2020). A deep learning based methodology for precipitation nowcasting with radar. *Earth and Space Science, 7*, e2019EA000812.

Chen, K., Liu, J., Guo, S., Chen, J., Liu, P., Qian, J., Chen, H., & Sun, B. (2016). Short-term precipitation occurrence prediction for strong convective weather using fy2-g satellite data: A case study of Shenzhen, South China. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 41*, 215.

Chhetri, M., Kumar, S., Pratim Roy, P., & Kim, B.-G. (2020). Deep BLSTM-GRU model for monthly rainfall prediction: A case study of Simtokha, Bhutan. *Remote Sensing, 12*(19), 3174.

Cristian, M. (2018). Average monthly rainfall forecast in Romania by using k-nearest neighbors regression. *Analele Universităţii Constantin Brâncuşi din Târgu Jiu: Seria Economie, 1*(4), 5–12.

Dacrema, M. F., Cremonesi, P., & Jannach, D. (2019). Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems* (pp. 101–109).

Damavandi, H. G., & Shah, R. (2019). A learning framework for an accurate prediction of rainfall rates. arXiv:1901.05885.

Dash, Y., Mishra, S. K., & Panigrahi, B. K. (2018). Rainfall prediction for the Kerala state of India using artificial intelligence approaches. *Computers & Electrical Engineering, 70*, 66–73.

Delleur, J. W., & Kavvas, M. L. (1978). Stochastic models for monthly rainfall forecasting and synthetic generation. *Journal of Applied Meteorology, 17*(10), 1528–1536.

Diez-Sierra, J., & del Jesus, M. (2020). Long-term rainfall prediction using atmospheric synoptic patterns in semi-arid climates with statistical and machine learning methods. *Journal of Hydrology, 586*, 124789.

Du, Y., Berndtsson, R., An, D., Zhang, L., Yuan, F., Uvo, C. B., & Hao, Z. (2019). Multi-space seasonal precipitation prediction model applied to the source region of the Yangtze river, China. *Water, 11*(12), 2440.

Du, J., Liu, Y., & Liu, Z. (2018). Study of precipitation forecast based on deep belief networks. *Algorithms, 11*(9), 132.

Du, J., Liu, Y., Yu, Y., & Yan, W. (2017). A prediction of precipitation data based on support vector machine and particle swarm optimization (PSO-SVM) algorithms. *Algorithms, 10*(2), 57.

Duong, T. A., Bui, M. D., & Rutschmann, P. (2018). A comparative study of three different models to predict monthly rainfall in Ca Mau, Vietnam. In *Wasserbau-Symposium Graz 2018. Wasserwirtschaft–Innovation aus Tradition. Tagungsband. Beiträge zum 19. Gemeinschafts-Symposium der Wasserbau-Institute TU München, TU Graz und ETH Zürich* (p. Paper–G5).

Gao, L., Wei, F., Yan, Z., Ma, J., & Xia, J. (2019). A study of objective prediction for summer precipitation patterns over eastern China based on a multinomial logistic regression model. *Atmosphere, 10*(4), 213.

Haidar, A., & Verma, B. (2018). Monthly rainfall forecasting using one-dimensional deep convolutional neural network. *IEEE Access, 6*, 69053–69063.

Htike, K. K., & Khalifa, O. O. (2010). Rainfall forecasting models using focused time-delay neural networks. In *International Conference on Computer and Communication Engineering (ICCCE'10)* (pp. 1–6). Piscataway: IEEE.

Huang, M., Lin, R., Huang, S., & Xing, T. (2017). A novel approach for precipitation forecast via improved k-nearest neighbor algorithm. *Advanced Engineering Informatics, 33*, 89–95.

Hung, N. Q., Babel, M. S., Weesakul, S., & Tripathi, N. K. (2009). An artificial neural network model for rainfall forecasting in Bangkok, Thailand. *Hydrology and Earth System Sciences, 13*(8), 1413–1425.

Hussein, E., Ghaziasgar, M., & Thron, C. (2020). Regional rainfall prediction using support vector machine classification of large-scale precipitation maps. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)* (pp. 1–8). Piscataway: IEEE.

Hussein, E. A., Ghaziasgar, M., Thron, C., Vaccari, M., & Bagula, A. (2021). Basic statistical estimation outperforms machine learning in monthly prediction of seasonal climatic parameters. *Atmosphere, 12*(5), 539.

Jing, J., Li, Q., & Peng, X. (2019). MLC-LSTM: Exploiting the spatiotemporal correlation between multi-level weather radar echoes for echo sequence extrapolation. *Sensors, 19*(18), 3988.

Karimi, H. A. (2014). *Big data: Techniques and technologies in geoinformatics*. Boca Raton: CRC Press.

Kashiwao, T., Nakayama, K., Ando, S., Ikeda, K., Lee, M., & Bahadori, A. (2017). A neural network-based local rainfall prediction system using meteorological data on the internet: A case study using data from the Japan meteorological agency. *Applied Soft Computing, 56*, 317–330.

Kumar, D., Singh, A., Samui, P., & Jha, R. K. (2019). Forecasting monthly precipitation using sequential modelling. *Hydrological Sciences Journal, 64*(6), 690–700.

Lakshmaiah, K., Murali Krishna, S., & Eswara Reddy, B. (2016). Application of referential ensemble learning techniques to predict the density of rainfall. In *2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT)* (pp. 233–237). Piscataway: IEEE.

Lee, J., Kim, C.-G., Lee, J. E., Kim, N. W., & Kim, H. (2018). Application of artificial neural networks to rainfall forecasting in the Geum river basin, Korea. *Water, 10*(10), 1448.

Lin, J. (2019). The neural hype and comparisons against weak baselines. In *ACM SIGIR forum* (vol. 52, pp. 40–51). New York: ACM.

Lu, J., Hu, W., & Zhang, X. (2018). Precipitation data assimilation system based on a neural network and case-based reasoning system. *Information, 9*(5), 106.

Ludewig, M., & Jannach, D. (2018). Evaluation of session-based recommendation algorithms. *User Modeling and User-Adapted Interaction, 28*(4–5), 331–390.

Mallika, M., & Nirmala, M. (2016). Chennai annual rainfall prediction using k-nearest neighbour technique. *International Journal of Pure and Applied Mathematics, 109*(8), 115–120.

Manandhar, S., Dev, S., Lee, Y. H., Meng, Y. S., & Winkler, S. (2019). A data-driven approach for accurate rainfall prediction. *IEEE Transactions on Geoscience and Remote Sensing, 57*(11), 9323–9331.

Mehdizadeh, S., Behmanesh, J., & Khalili, K. (2018). New approaches for estimation of monthly rainfall based on GEP-ARCH and ANN-ARCH hybrid models. *Water Resources Management, 32*(2), 527–545.

Mehr, A. D., Nourani, V., Khosrowshahi, V. K., & Ghorbani, M. A. (2019). A hybrid support vector regression–firefly model for monthly rainfall forecasting. *International Journal of Environmental Science and Technology, 16*(1), 335–346.

Mishra, N., & Kushwaha, A. (2019). Rainfall prediction using gaussian process regression classifier. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 8*(8), 392–397.

Mohamadi, S., Ehteram, M., & El-Shafie, A. (2020). Accuracy enhancement for monthly evaporation predicting model utilizing evolutionary machine learning methods. *International Journal of Environmental Science and Technology, 17*, 1–24.

Mosavi, A., Ozturk, P., & Chau, K.-W. (2018). Flood prediction using machine learning models: Literature review. *Water, 10*(11), 1536.

Nasseri, M., Asghari, K., & Abedini, M. J. (2008). Optimized scenario for rainfall forecasting using genetic algorithm coupled with artificial neural network. *Expert Systems with Applications, 35*(3), 1415–1421.

Nielsen, A. (2020). *Practical time series analysis: Prediction with statistics and machine learning*. Sebastopol: O'Reilly.

Nourani, V., Uzelaltinbulat, S., Sadikoglu, F., & Behfar, N. (2019). Artificial intelligence based ensemble modeling for multi-station prediction of precipitation. *Atmosphere, 10*(2):80.

Oswal, N. (2019). Predicting rainfall using machine learning techniques. arXiv:1910.13827.

Pan, B., Hsu, K., AghaKouchak, A., & Sorooshian, S. (2019). Improving precipitation estimation using convolutional neural network. *Water Resources Research, 55*(3), 2301–2321.

Pantanowitz, A., & Marwala, T. (2009). Missing data imputation through the use of the random forest algorithm. In *Advances in Computational Intelligence* (pp. 53–62). Berlin: Springer.

Parmar, A., Mistree, K., & Sompura, M. (2017). Machine learning techniques for rainfall prediction: A review. In *International Conference on Innovations in Information Embedded and Communication Systems*.

Patel, M., Patel, A., Ghosh, R. (2018). Precipitation nowcasting: Leveraging bidirectional LSTM and 1d CNN. arXiv:1810.10485.

Peng, Y., Zhao, H., Zhang, H., Li, W., Qin, X., Liao, J., Liu, Z., Li, J. (2019). An extreme learning machine and gene expression programming-based hybrid model for daily precipitation prediction. *International Journal of Computational Intelligence Systems, 12*(2), 1512–1525.

Pham, Q. B., Abba, S. I., Usman, A. G., Linh, N. T. T., Gupta, V., Malik, A., Costache, R., Vo, N. D., & Tri, D. Q. (2019). Potential of hybrid data-intelligence algorithms for multi-station modelling of rainfall. *Water Resources Management, 33*(15), 5067–5087.

Ramsundram, N., Sathya, S., & Karthikeyan, S. (2016). Comparison of decision tree based rainfall prediction model with data driven model considering climatic variables. *Irrigation and Drainage Systems Engineering, 5*(3).

Sardeshpande, K. D., & Thool, V. R. (2019). Rainfall prediction: A comparative study of neural network architectures. In *Emerging Technologies in Data Mining and Information Security* (pp. 19–28). Berlin: Springer.

Sato, R., Kashima, H., & Yamamoto, T. (2018). Short-term precipitation prediction with skip-connected PredNET. In *International Conference on Artificial Neural Networks* (pp. 373–382). Berlin: Springer.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using mice: A caliber study. *American Journal of Epidemiology, 179*(6), 764–774.

Shenify, M., Danesh, A. S., Gocić, M., Taher, R. S., Wahab, Ainuddin, W. A., Gani, A., Shamshirband, S., & Petković, D. (2016). Precipitation estimation using support vector machine with discrete wavelet transform. *Water Resources Management, 30*(2), 641–652.

Shi, E., Li, Q., Gu, D., & Zhao, Z. (2017). Convolutional neural networks applied on weather radar echo extrapolation. In *DEStech Transactions on Computer Science and Engineering* (case), 695–704. DEStech Publications.

Shi, X., Chen, Z., Wang, H., Yeung, D., Wong, W., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. ArXiv, abs/1506.04214.

Shi, X., Gao, Z., Lausen, L., Wang, H., Yeung, D.-Y., Wong, W., Woo, W.-C. (2017). Deep learning for precipitation nowcasting: A benchmark and a new model . In *Advances in Neural Information Processing Systems* (pp. 5617–5627).

Shi, X., & Yeung, D.-Y. (2018). Machine learning for spatiotemporal sequence forecasting: A survey. arXiv:1808.06865.

Singh, G., & Kumar, D. (2019). Hybrid prediction models for rainfall forecasting. In *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 392–396). Piscataway: IEEE.

Singh, S., Sarkar, S., & Mitra, P. (2017). Leveraging convolutions in recurrent neural networks for doppler weather radar echo prediction. In *International Symposium on Neural Networks* (pp. 310–317). Berlin: Springer.

Sulaiman, J., & Wahab, S. H. (2018). Heavy rainfall forecasting model using artificial neural network for flood prone area. In *IT Convergence and Security 2017* (pp. 68–76). Berlin: Springer.

Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 10*(6), 363–377.

Tran, Q.-K., & Song, S.-K. (2019a). Computer vision in precipitation nowcasting: Applying image quality assessment metrics for training deep neural networks. *Atmosphere, 10*(5), 244.

Tran, Q.-K., & Song, S.-K. (2019b). Multi-channel weather radar echo extrapolation with convolutional recurrent neural networks. *Remote Sensing, 11*(19), 2303.

Valencia-Payan, C., & Corrales, J. C. (2018). A rainfall prediction tool for sustainable agriculture using random forest. In *Mexican International Conference on Artificial Intelligence* (pp. 315–326). Berlin: Springer.

Wang, Y., Long, M., Wang, J., Gao, Z., & Philip, S. Y. (2017). PredRNN: Recurrent neural networks for predictive learning using spatiotemporal LSTMs. In *Advances in Neural Information Processing Systems* (pp. 879–888).

Weesakul, U., Kaewprapha, P., Boonyuen, K., & Mark, O. (2018). Deep learning neural network: A machine learning approach for monthly rainfall forecast, case study in eastern region of Thailand. *Engineering and Applied Science Research, 45*(3), 203–211.

Xu, L., Chen, N., Zhang, X., & Chen, Z. (2020). A data-driven multi-model ensemble for deterministic and probabilistic precipitation forecasting at seasonal scale. *Climate Dynamics, 54*, 3355–3374.

Yu, P.-S., Yang, T.-C., Chen, S.-Y., Kuo, C.-M., & Tseng, H.-W. (2017). Comparison of random forests and support vector machine for real-time radar-derived rainfall forecasting. *Journal of Hydrology, 552*, 92–104.

Zainudin, S., Jasim, D. S., & Bakar, A. A. (2016). Comparative analysis of data mining techniques for Malaysian rainfall prediction. *International Journal on Advanced Science, Engineering and Information Technology, 6*(6), 1148–1153.

Zhan, C., Wu, F., Wu, Z., & Chi, K. T. (2019). Daily rainfall data construction and application to weather prediction. In *2019 IEEE International Symposium on Circuits and Systems (ISCAS)* (pp. 1–5). Piscataway: IEEE.

Zhang, C.-J., Zeng, J., Wang, H.-Y., Ma, L.-M., & Chu, H. (2020). Correction model for rainfall forecasts using the LSTM with multiple meteorological factors. *Meteorological Applications, 27*(1), e1852.

Zhang, P., Jia, Y., Gao, J., Song, W., & Leung, H. K. N. (2018). Short-term rainfall forecasting using multi-layer perceptron. *IEEE Transactions on Big Data, 6*, 93–106.

Zhuang, W. Y., & Ding, W. (2016). Long-lead prediction of extreme precipitation cluster via a spatiotemporal convolutional neural network. In *Proceedings of the 6th International Workshop on Climate Informatics: CI.*