Evaluation of ECMWF’s forecasting system for probabilistic urban flood prediction: a case study in Mexico City

Marco R. López a,*, Adrián Pedrozo-Acuña b and Marcela L. Severiano Covarrubias a

a Hydraulic Department, Engineering Institute UNAM, Mexico City 04510, México
b Instituto Mexicano de Tecnología del Agua IMTA, Jiutepec 62550, México
*Corresponding author. E-mail: mrodrigo.lz18@gmail.com

ABSTRACT

As the world continues urbanizing, including efforts to forge a new framework of urban development is necessary. Recent studies related to flood prediction and mitigation have shown that Ensemble Prediction Systems (EPSs) constitute a valuable and essential tool for an Early Warning System. However, the use of EPS for flood forecasting in urban zones has yet to be understood. This work has the objective to investigate the potential use of the Operational EPS, issued by the European Centre for Medium-Range Weather Forecasts (ECMWF), for probabilistic urban flood prediction. In this research, a precipitation forecast verification was carried out in two study zones: (1) Mexico Valley Basin and (2) Mexico City, where for the latter, forecasts were compared against real-time observed data. The results showed good forecast reliability for a rain threshold of up to 20 mm in 24-hourly accumulations, with the first 36 h of the forecast horizon being the most reliable. The EPS has sufficient resolution and precision for flood prediction in Mexico City, which represents a further step toward developing a flood warning system at the local level based on ensemble forecasts.

Key words: ensemble forecast, flood forecasting, forecast verification, probabilistic forecast, urban floods

HIGHLIGHTS

- ECMWF’s forecast evaluation from real-time measurements and at a temporal resolution of less than 12 h.
- Implementation of a predictive model to predict the occurrence of a flood event.
- To extend the use of EPS to include them in the chain for emergency and decision-making.
- This work seeks to shorten the gap of previous research and direct its efforts to the evaluation of urban and flash floods.

INTRODUCTION

Urbanization brings together various opportunities for economic development and the well-being of society. However, with the high population density, the concentration of assets and more than 80% of the cities located in watersheds, floods represent a major challenge for their sustainability in the future. The United Nations estimates that the world’s population living in urban areas will have a global increase of 68% by 2050 (United Nations 2018). Most of this growth will occur in developing countries, mainly in megacities such as Mexico City, as well as in medium-sized cities, with a consequent increase in the number of people exposed to floods.

Traditional flood management is based on protection using structural measures to reduce peak flow and flood extension; however, such measures fail to completely eliminate risk (Jain et al. 2018). These measurements are even more limited due to the uncertainty in magnitude, lead time, geographic extension and geophysical interactions of floods (Moore et al. 2005). For all these reasons, forecasting is fast becoming a key instrument to respond to the demand for better risk management due to the exposure of infrastructure and people to floods and is associated with the uncertainty of future events (Krzysztofowicz 2001).

The most common forecasting systems incorporate quantitative precipitation forecasts (e.g. René et al. 2013) either from radar data, Numerical Weather Prediction (NWP) or a combination of both. As a result, from NWP, the generation of Ensemble Prediction Systems (EPSs) represents the attempt to estimate the range of possible scenarios based on a set of initial.
conditions and to counteract the limitations associated with deterministic models (Jain et al. 2018). Systems such as the European Flood Alert System (Thielen et al. 2008) and the Global Flood Awareness System (Alfieri et al. 2013) are examples of how EPSs have been used for flood prediction at resolutions greater than 30 km and for fluvial flooding warning systems. In Mexico, one of the first implementations of an EPS was documented by Rodríguez-Rincón et al. (2015), who used a Numerical Weather Prediction Model in combination with a distributed hydrological model for the generation of a runoff ensemble, on Tonalá river, Tabasco.

Several researches have demonstrated that probabilistic flood prediction can play a crucial role in ensuring a more conscious decision-making and considering the uncertainty of meteorological phenomena, both globally and locally (e.g. Moore et al. 2005; Verbunt et al. 2007; Pedrozo-Acuña et al. 2013; Revilla-Romero et al. 2015; Emerton et al. 2016; Lee et al. 2018). Among the new research methodologies and techniques for flood risk reduction and mitigation based on EPS are those used by the European Centre for Medium-Range Weather Forecasts (ECMWF, https://www.ecmwf.int/) and the Joint Research Centre (JRC), which consists of forecasting systems for climate variables at the global scale, in order to provide information in an emergency situation to grant an adequate response time and also an effective tool for forecasters and decision makers. Previous research on ECMWF data agrees that a probabilistic ensemble-based forecast always performs better than a single forecast for any lead time (Roulin 2006; Verbunt et al. 2007) and that the greatest benefits are passed on to the user based on the choice of appropriate probability thresholds for decision-making.

This work is aimed at creating an architecture for urban flood forecasting described by Henonin et al. (2013) and is composed of four main components: (1) inputs, (2) data collection, (3) modeling and decision support tools and (4) warning process.

Based on the first component, and due to the uncertainty in flood forecasting, it is necessary to characterize and evaluate the performance and quality of forecasts. To do this, different attributes are examined, such as reliability, resolution, discrimination and sharpness of a prediction (e.g. Murphy 1993; Buizza et al. 2000). This process is known as Forecast Verification (Jolliffe & Stephenson 2013). However, most studies on flood prediction and forecast verification have been limited to dealing only with fluvial flooding (Beg et al. 2019; Habibi et al. 2019) and have been carried out at continental and basin scale with a temporal resolution greater than or equal to 12 h. Therefore, it is necessary to address the use of EPS for nonfluvial flood forecasting, as in the case of Mexico City, where floods are mainly caused by convective rains and there are other considerations to take into account, such as the presence of structures and drainage.

The current challenges in flood forecasting should extend the use of ensemble forecasts for operational systems and include them in the chain for emergency decision-making and water management. In summary, it is important that current and future work on flood forecasting target local-scale flood events that are related to urban or flash floods (Wu et al. 2020). Mexico City is the perfect example to test the use of EPS, because the city is completely dependent on the sewer system to prevent and mitigate flooding. For these reasons, prevention strategies to mitigate current and future floods in the city require a correct estimation of the atmospheric conditions that could arise in the future.

The aim of this study is to investigate the potential of ECMWF’s ensemble forecasts as a tool for urban flood prediction. This paper seeks to shorten the limitations of previous research by evaluating ECMWF’s forecasts at a temporal resolution of less than 12 h and provides an important opportunity to advance in the development of a flood forecasting system for Mexico City. This study seeks to address the following research questions: how well the forecasts correspond to observed precipitation in the Valley of Mexico and what is the accuracy and reliability of the forecasts and whether its resolution is sufficient for flood prediction in the city. To do this, a forecast verification was carried out for two study zones: (1) Mexico Valley Basin (MVB) and (2) Mexico City (CDMX). For zone 1, the forecasts were verified against 24-hourly observed accumulation issued by the CLICOM system (clicom-mex.cicese.mx), over 2007–2014. For zone 2, forecasts were verified against observed real-time data issued by the Hydrological Observatory of Engineering Institute of UNAM (OHIIUNAM, https://oh-iiunam.mx/), over the rainy season (MJJASON, May–November) of 2017–2019. Verification was carried out in terms of the following scalar attributes: precision, reliability, resolution, discrimination and performance. The ensemble probabilities were estimated by a predictive model for the subsequent application of probabilistic verification methods.

This work is organized as follows. The section ‘Object of investigation’ describes the location and main characteristics of the study area. The section ‘Methods’ describes the datasets and verification framework. The section ‘Results and discussion’ presents the verification results in MVB; findings of 6-hourly accumulated precipitation verification forecasts in Mexico City and discussion of the results; and finally, conclusions.
OBJECT OF INVESTIGATION

Where Mexico City now sits, there were five lakes located in the lower part of a hydrological basin covering an area of 1,500 km². The relationship of the city with water changed radically with the arrival of the Spanish conquerors in 1521. This started the elimination of the lake culture and the beginning of the drainage of water from MVB, with the disappearance of about 1,100 km² of water along 400 years.

México City (Figure 1(a), zone 2) is located more than 2,000 meters above sea level within the sub-basin of Lake of Texcoco and Zumpango; belonging to the Valley of Mexico Basin (Figure 1(a), zone 1) in Hydrological Region 26. The difference in heights within the Valley causes it to occur from a humid climate in the mountainous area, to a dry and hot climate in the lower areas. The average annual precipitation varies in a range from 600 to 1,500 mm, generally distributed from May to November (MJJASON).

Throughout its history, the Valley of Mexico has suffered several impacts due to drainage of the basin and the overexploitation of the aquifer, problems that stem directly from the exacerbated urban growth. This growth has other environmental implications within the Valley; for example, asphalt coating and concrete have eaten up green areas (Connolly 1999), creating an impermeable layer. This prevents natural infiltration into the aquifers and, consequently, increases the load on the drainage system. Together, these factors have generated an alteration of the hydrological cycle within the basin that represents a present and future risk for city population.

During the rainy season, floods and puddles are incidents that are repeated every year (Figure 1(b)). For example, in 2017, there were 738 emergency reports (Figure 1(b)) due to the occurrence of meteorological phenomena and attended by the Government of Mexico City (Secretaría de Gestión Integral de Riesgos y Protección Civil 2019).

METHODS

Data

Forecast data

Forecast data used in this work correspond to those emitted by the ECMWF global atmospheric model, since they offer the best spatial and temporal resolution required by the zone and the studied phenomenon. ECMWF is a research institute with a
service that operates 24 h a week. Since 1992, it produces meteorological ensembles and analyses that describe possible scenarios and their probability of occurrence at the global scale, with forecast horizons ranging from a medium, monthly and temporal range, up to a year ahead. ECMWF offers high-frequency products (hourly products) for the first 90 h and issued four times a day. The access to ECMWF data is carried out through its Meteorological Archival Retrieval System. The files are stored in formats prescribed by the World Meteorological Organization (WMO), called GRIB (for field forecasts), and all parameters are available in a grid point space (latitude-longitude). For a more detailed description of ECMWF datasets, consult *User Guide to ECMWF Forecast Products* (Owens & Hewson 2018).

We gathered data from ENS (Ensemble Forecast) and HRES (High-Resolution Forecast) products. The parameter used was total precipitation (tp, 228.18) at the surface level (sf), for 12:00 UTC (base time) and a lead time of 90 h (lt = 0 + 90 h). This parameter consists of the sum of Convective Precipitation (CP) and Large-Scale Precipitation (LSP). The selection of lead time was established based on the type of product, its spatial and temporal resolution and an extension of the studied area (see, Emerton et al. 2016).

ENS provides an estimate of the reliability of a single forecast. It consists of 50 members with disturbed initial conditions (Persson 2015) with a resolution of 0.125° plus a Control Member (CF) without disturbed initial conditions. Each member of the ensemble constitutes an individual deterministic forecast; then, this set of deterministic forecasts is considered as an estimate of the distribution of a variable x, which varies for each point within a grid and at each time step of the forecast horizon (Wilson et al. 1999). In contrast, HRES is a unique or nonprobabilistic prediction (deterministic forecast) and currently constitutes the model with the best resolution (0.1°) of ECMWF. HRES provides a more detailed description of future climate in comparison with the Control Forecast (CF) or with any individual member of the ensemble.

Previous research has shown that there is a close relationship between both ENS and HRES forecasts, and it is advisable to use a combination of both (Owens & Hewson 2018). For example, Richardson et al. (2015) found that the nonprobabilistic HRES forecast can have a probability weight greater than 40% in the first 24 h of the forecast, and for this reason, it is advisable to evaluate the behavior of both products, and, thus, be able to improve the predictions.

**Observed data**

For forecast verification in study zone 1, we collect daily data from 103 weather stations (CLICOM System, clicom-mex.cicese.mx). The period of analysis and selection of stations was established based on quality, data continuity, spatial coverage and availability of ECMWF forecast data, resulting in a period of 8 years of daily observed precipitation (2007–2014). The motivation for this was to have enough historical data to obtain robust results of forecast performance.

On the other hand, to evaluate the forecasts in study zone 2 (CDMX), we used observed real-time data for the rainy season (MJJASON) over 2017–2019, issued by OH-IIUNAM. In mid-2015, the Engineering Institute of the National Autonomous University of Mexico (UNAM) launched the hydrological observatory program to more accurately measure precipitation in the city. By 2017, the system consisted of 10 measuring stations, and by May 2019, the number of stations increased to 55 (yellow dots in Figure 1). This system is based on low-cost devices for data acquisition to be transferred to a central server and, thus, be available through a web platform (https://www.oh-iiunam.mx/). The individual stations operate independently and are composed of a laser-optical disdrometer (Parsivel2, manufactured by OTT Hydromet) to measure precipitation on 1-min time resolution with an accuracy of ±5% for liquid precipitation and ±20% for solid precipitation (Pedrozo-Acuña et al. 2017). Table 1 presents a summary of datasets used in this work.

**Forecast verification framework**

A good ensemble forecast should have the following three qualities:

1. The probability distribution must be precise.
2. The forecasts must correspond well for events of different magnitudes.
3. The probability of the forecasts must be consistent with the observed frequencies.

To evaluate the above features, it is necessary to carry out a forecast verification process (see Jolliffe & Stephenson 2013 and Chapter 4.4).

Forecast verification is a crucial stage in flood forecasting and there is a wide variety of verification methods (see Jolliffe & Stephenson 2013; Brooks et al. 2015), all of which involve the relationship between forecasts and observations of a given event. A common process in verification is to compare the magnitude and location of precipitation and flood depth by applying numerous performance metrics (e.g. Hamill et al. 2008; Jolliffe & Stephenson 2013; Li et al. 2021). The application of
Verification methods depend on the type of forecast, whether probabilistic or nonprobabilistic, as well as the characteristics of the study area, whether it is local (punctual) or a field forecast (Wilks 2006).

In this work, forecast verification was carried out ex post (see Jolliffe & Stephenson 2013, p. 10). We used a variety of nonprobabilistic and probabilistic verification methods to evaluate HRES and ENS quality, respectively, in terms of the scalar attributes: precision, reliability, resolution, discrimination and performance (see Murphy 1993).

To obtain a numerical value of the above attributes, we used the following verification methods: eyeball, BIAS, root mean squared error (RMSE), ensemble spread, reliability diagram and receiver operating characteristic (ROC) curves.

Prior to the verification process, ENS data were linearly interpolated to 0.1° resolution to be consistent with the HRES grid. Verification metrics were computed using the forecast grid cells, and in order to compare forecasts against observed data, weather stations (point locations) within a grid cell were arithmetically averaged.

To apply BIAS, RMSE and meteogram plot (see Figure 9), we used the Verif tool in its version 1.1.0 (Nipen et al. 2019). This tool allows the application of most of the metrics described by Brooks et al. (2015) for point locations and produces graphical outputs for the verification methods. Verif works with a data-based input text file with forecasts and observations according to a specific format. This data-based input must include information on dates, time horizons and locations, so that statistics can be calculated for different dimensions of the forecast (e.g. time window, lead time, date and location).

Forecast data were processed using a Python interface called cfgrib (https://github.com/ecmwf/cfgrib) that allows to read GRIB files with xarray Python package (http://xarray.pydata.org/en/stable/api.html). Processing of the information involved data cleaning, as well as the generation of dummy variables (a simpler variable that usually takes values of 0/1) for the categorization of events. The objective is to restructure the forecast files and datasets into a table or data structure, which in the case of Python, is represented by a Data Frame. Null values were removed from the data-based input in order to avoid incorrect behaviors between variables and erroneous interpretations of information.

Finally, to evaluate ensemble forecasts, it is necessary to convert them into probabilistic information. To do this, we estimated ensemble probabilities using a Logit Model implemented with Python’s library Scikit-learn (Pedregosa et al. 2011; https://scikit-learn.org).

For the verification process, two study areas were considered: For zone 1, the 24-h forecast of HRES and ENS was verified over 2007–2014. For zone 2, verification was conducted using the mean and 95th percentile of accumulated rain in 6 h, for a lead time of 0–90 h (lt = 0 + 90 h) over the rainy season (MJJASON) of 2017–2019.

Logistic regression model

The simplest method to estimate the ensemble probabilities is by means of the cumulative distribution function, where the probability is given by the number of members who predict an event divided by the total number of members of the ensemble (e.g. Buizza et al. 2000; Atger 2001); nevertheless, this methodology hardly returns calibrated probabilities, and consequently, does not reflect the true occurrence of the event in question. In recent years, predictive models have been implemented for estimating and calibrating the ensemble forecast probabilities (e.g. Niculescu-Mizil & Caruana 2005; Gweon & Yu 2019).

Predictive models consist of a set of statistical algorithms that, when applied to historical data, return a mathematical function to solve a specific problem. In this case, the problem is to predict a value in terms of probability with values between 0 and 1, with the objective of using the estimated probabilities to predict the occurrence of a given event.

When estimating the probabilities of an ensemble, it is desired that these reflect the true frequency of the event of interest (Kuhn & Johnson 2013). In other words, the predicted class (refers to the event or quantity that is predicted) probability must

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Summary of forecast products used in this study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study zone</td>
<td>Dataset</td>
</tr>
<tr>
<td>MVB</td>
<td>HRES</td>
</tr>
<tr>
<td></td>
<td>ENS</td>
</tr>
<tr>
<td></td>
<td>CLICOM</td>
</tr>
<tr>
<td>CDMX</td>
<td>HRES</td>
</tr>
<tr>
<td></td>
<td>CF</td>
</tr>
<tr>
<td></td>
<td>ENS</td>
</tr>
<tr>
<td></td>
<td>OH-HIUNAM</td>
</tr>
</tbody>
</table>
be well calibrated. To do this, we used a Logit Model (Logistic Regression) implemented with Python's library Scikit-learn (https://scikit-learn.org).

Logistic regression is a variant of linear regression where the output variable is categorical or classified with values of 0 and 1 and has proven useful for predicting well-calibrated probabilities based on historical data (e.g. Hamill et al. 2008; Bartosz et al. 2018; Browniee 2018). This model is linear in its parameters, which are estimated by maximum likelihood, and it is defined by the following equation:

\[
P = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)]}
\]

where \( p \) is the event probability, \( P \) is the number of predictors, \( \beta \) is the model parameter and \( x \) is the predictor variable. For a more detailed description of the logistic regression model, consult Kuhn & Johnson (2013, p. 282).

To avoid deviations or biases due to the origin of the information, the data-based input was randomly divided into two sub-sets: 70% of the data for training and 30% for testing. HRES forecast was included as another member of the ensemble, since the combination of ENS and HRES was considered to have an added value (see Gouweleeuw et al. 2005).

**Reliability diagram**

The most widely used method to evaluate the reliability, resolution and precision of a probabilistic forecast is the reliability diagram (Bröcker & Smith 2007). Reliability curve or calibration plot represents a method to assess the quality of the class probabilities. This diagram answers how well the probabilities of an event correspond to its observed frequencies. The position of the curve with respect to the 45° diagonal helps to interpret the results. If the curve is located below the diagonal, it means that the model or system is over-forecasting; if the curve is located above the diagonal, it means that the model or system under-forecasts the events. The better calibrated or more reliable a forecast, the closer the curve will appear along the main diagonal from the bottom left to the top right of the plot.

**ROC curve**

ROC curves have been used increasingly in recent years to evaluate probability forecasts for binary predictions (yes/no). ROC is a plot of hit rate (POD) versus false alarm rate (POFD) using increasing probability thresholds and allows to measure the forecast ability to discriminate observed and unobserved events. Hit rate or probability of detection (POD) provides a measure of forecast discrimination ability and indicates the fraction of observed events predicted correctly with a perfect value equaling to 1. False alarm rate, also called probability of false detection (PODF), is the ratio of false alarms to the total number of nonoccurrences of the event, with a perfect value equal to 0. The shape of the ROC curve is determined by the intrinsic discrimination capacity of the forecasting system and does not provide any information on reliability (Jolliffe & Stephenson 2013). This metric is conditioned by the observed events, that is, given that an event occurred, what was its corresponding forecast (Brooks et al. 2015).

ROC curve can be used to quantitatively evaluate the performance of the forecasting model. The perfect forecasts or those with better discrimination exhibit a curve closer to the upper left corner of the diagram and its area under the curve (AUC) will be equal to 1, whereas, for an ineffective model or system, it results in a curve close to a 45° diagonal with the AUC equal to 0.50. In common practice, it is considered that an area under the ROC curve greater than 0.80 is indicative of a good forecasting system, and the minimum limit for a useful system is for AUC equal to or greater than 0.70 (Buizza et al. 2000). In this work, we used Python's library Scikit-learn (Pedregosa et al. 2011, https://scikit-learn.org) to implement ROC curves.

**RESULTS AND DISCUSSION**

**Forecast verification in MVB**

To assess the correspondence between forecasts and observations of 24-hourly precipitation accumulation and to compare the difference between ENS and HRES in relation to different ranges of rain, both low intensity and extreme events, the results of forecast verification in the Valley of Mexico basin are summarized below.
Preliminary exploration of forecasts – eyeball, BIAS and RMSE

Before converting the forecasts to probabilistic information, their basic statistical properties were analyzed. In a first phase, the verification focused on HRES and the ensemble mean (EMEAN). The verification metrics applied in this case were eyeball, BIAS and RMSE.

The limitation of the Eyeball method is the loss of temporal information, so to counter this, Figure 2 shows a visual comparison between CLICOM, HRES and EMEAN for the 24-h average precipitation per month in MVB. To interpolate the data within the basin, we used an inverse weighting distance interpolation. This comparison shows a good consistency of forecasts month by month. What stands out in this map is that the months with the highest rainfall (MJJASON) are properly identified by the forecasts. On the other hand, EMEAN tends to over-forecast rain, especially small events located in the range of 0–3 mm in 24 h.

In order to assess the average magnitude of forecasts errors, Figure 3 presents the averaged BIAS and RMSE in the basin for HRES, EMEAN and CLICOM. Bias is a comparison of the average forecast with the average observation and it is a measure of discrimination capacity of the forecast. As shown in Figure 3(a), BIAS is greater for the rainy season (MJJASON), indicating over-forecast when BIAS>0 and under-forecast when BIAS<0.

In general, there is a tendency to over-forecast events. The BIAS of both HRES and EMEAN is similar, although EMEAN has better performance in comparison. Turning now to Figure 3(b), the results show a year-to-year pattern, where the greatest errors occur over the rainy season (MJJASON). According to the figure, in this season, in average, the error varies between 0 and 20 mm.

Probabilistic verification – ensemble spread, reliability and ROC

Spread is an indirect measure of forecast accuracy and is given by the standard deviation of the ensembles. If the ensemble distribution is appropriate, the observed value will have the same probability of occurrence at any percentile of the distribution (Wilks 2006, p. 234). Figure 4 displays a map with the ensemble spread and the observed precipitation. According to the isohyets, the observed precipitation varies in an average range from 0 to 4 mm in 24 h. As can be seen in Figure 4, spread is greater in the upper areas of the basin, with an average variation of less than 0.5 mm.

Figure 5 shows the reliability diagram and ROC curves for probabilistic forecasts of 24-hourly accumulated precipitation in MVB. Three thresholds were established: 5, 10 and 20 mm of accumulated rain in 24 h, to consider the occurrence of scarce to heavy rains according to CENAPRED (2014). For this analysis, the event was considered to occur if the forecasted probability was greater than 0.5.

Figure 6(a) shows a higher reliability for thresholds from 0 to 10 mm. For a precipitation threshold less than 5 mm, there is under-forecasting of events associated with probabilities less than 50%, which may be due to the fact that, in general, ENS predicts higher values for light rain events. On the other hand, for rains greater than 5 mm, there is an over-forecast of events that have a high probability of occurrence (P>60%). For rains greater than 10 mm, the forecast is biased and there is an over-forecast for high probability events. Finally, for a threshold of 20 mm, the shape of the curve indicates that it is probably under-sampled (e.g. Ebert 2007) and, hence, is not representative of the event. In this case, it would be necessary to have more historical data to obtain a good estimate of the probabilities of this threshold.

ROC curves in Figure 5(b) indicate that forecasts present a good discrimination ability. Furthermore, the AUC can be considered as another verification metric. Interestingly, AUCs for all analyzed thresholds turned out to be around 0.85. This could be due to the fact that the model fits well to true negatives for all thresholds, for which the frequency is very high. This indicates a low probability of false alarms, which contributes to the discrimination capacity of the forecast.

Forecast verification in Mexico City (CDMX)

For Mexico City, it is considered a flood when the water depth exceeds 20 cm. This is a height that is considered sufficient to cause damage to homes and generate problems and obstructions in the transport routes (Reinooso et al. 2012).

The runoff that occurs in the areas of hills and mountains, such as torrential rains where precipitation is greater than the drainage capacity, are the main factors that give rise to floods in the city. The areas that have a low level of vulnerability are those corresponding to the mountainous region, formed by volcanic rocks. Despite this, the rains that occur in the higher areas produce high-speed runoff, causing flash floods in the lower areas that are more susceptible to flooding (Figure 6).

The rains generally move from the northeast to the southwest of the city (León & Neri 2010). Convective rains are the main source of floods in Mexico City, that is why the spatial distribution of precipitation plays a key role. For this reason, it is
Figure 2 | Eyeball maps: 24-hourly accumulated precipitation per month in the Valley of Mexico Basin, period 2007–2014.
Figure 3 | BIAS and RMSE in the Valley of Mexico Basin for the 24-h rain forecast given by HRES and EMEAN.

Figure 4 | Average ensemble spread of accumulated rain in 24 h for CVM, period 2007–2014.
important to analyze precipitation forecasts on a smaller timescale; therefore, the accumulated precipitation every 6 h was assessed considering a lead time of 90 h (lt = 0 + 90 h). Knowing the forecast for a 6-h horizon can be very useful for emergency services, since they can focus their efforts on strategic points in the city with a sufficient amount of time.

**Flood events and meteogram**

In this first analysis, the precision and reliability of the forecasts in the grid cells within the study area was evaluated, for which two flood events were considered and using a time base in advance of 2 days: June 29, 2017 (Aristegui 2017) and August 20, 2018 (López 2018). CDMX was divided into 14 zones according to the HRES grid (Figure 6).

**Figure 5** | Reliability and ROC curves for the 24-h forecast in the Valley of Mexico Basin.

**Figure 6** | Map of flood risk in CDMX and flood events of June 29, 2017 and August 20, 2018.
During the occurrence of the event of June 29, 2017 (Figure 7), an extraordinary rain was presented, which reached 54 mm; which corresponds to rainfall associated with a return period of 10 years in the area (zone 5 in Figure 6). Since the afternoon of June 28, 2017, a heavy rain fell on Mexico City. Polanco was the most affected area along with ‘Circuito Interior’ avenue (one of the main roads of the city), where the flood exceeded 1-m height (location a in Figure 6). On the other hand, ‘Indios Verdes’ zone (location b in Figure 6) was also affected and it was necessary to rescue people who were stranded inside their vehicles (EL PAÍS 2017; EL UNIVERSAL 2017; LA SILLA ROTA 2018; Redacción Pásala 2018).

**Figure 7** | Event June 29, 2017: analysis of the values per grid cell of the HRES and ENS forecast and maps showing the accumulated rainfall results for a horizon of 90 h. P95, 95th percentile; P05, 5th percentile.
According to Figure 7, the observed event corresponds adequately to that estimated by the ensembles. In zones 1, 3, 4, 6, 9 and 10, the event is located within the distribution of the ensembles between percentiles 5 and 95 curves. After the first 12 h, HRES underestimates the precipitation in zones 4, 5, 9 and 10 (Figure 7), which are the zones where it rained the most, exceeding 40 mm. This indicates that in the presence of greater precipitation and as the lead time increases, HRES loses reliability, which agrees with the study carried out by Richardson et al. (2015). In all zones, the CF behaves very similarly to the mean of the ensembles. Zones 3 and 5 are where the greatest impacts occurred, and according to the risk map in Figure 6, these zones correspond to the highest risk of flooding. In zone 5 plot, it is observed that the event was not correctly captured by ENS due to the rapid evolution of atmospheric conditions; however, the 95th percentile of ENS manages to resemble the observed event (map e in Figure 7), which indicates that some members of the ensemble may be representative of extreme events.

For the event of August 20, 2018 (Figure 8), a yellow alert was activated (15–29 mm/24 h, Secretaría de Gestión Integral de Riesgos y Protección Civil 2021) due to the displacement of the storm over 13 of the city mayors and a red alert (50–70 mm/24 h) in Cuajimalpa (location c in Figure 6), which corresponds to a return period in the area of 10 years. This caused a severe flood in ‘Tolua’ avenue, and the security protocols for the subway were activated (Pérez Caballero 2018).

In Figure 8, it is observed that for zones 1, 6, 10, 12 and 13, HRES overestimates the event. As in the previous event, it is observed that the first 36 h are the most reliable, and the similarity between EMEAN and the CF is visible again in all the zones.

As in the event of June 29, 2017, in this case, the HRES forecast also underestimates the events of greater magnitude, which is observed for zones 4, 8 and 9. What is interesting about this case is that ENS was able to identify the rapid change in the observed precipitation for zone 4. Zone 4 is an upper part of the city where there are steep slopes, so even though there were no floods, runoff from the streets and avenues also affected the area. On the other hand, in zone 11 (location d in Figure 6), although the rain alert only reached yellow level (15–29 mm/24 h); being one of the lowest areas of the city, it is more vulnerable and with a greater probability of flooding. This last observation can open the doors so that the alert scale is not only a function of the magnitude of the rain, but also of the vulnerability of the area and the conditions of the terrain.

To complement the above, the meteogram in Figure 9 shows the average evolution of observed and forecasted precipitation over the rainy season (MJJASON) in Mexico City. The red and blue lines correspond to the average evolution of the precipitation issued by OHIUNAM and HRES for a lead time of 90 h, respectively. The enveloping lines correspond to the percentiles of the probabilistic forecast (ENS) calculated for a range of 10–90%. The results indicate that observed precipitation falls within the 20–30% percentile of the probability distribution and generally below the HRES curve (red line), which is indicative of over-forecasting. On the other hand, the forecast is quite accurate for a lead time of 36 h.

**Probabilistic verification – reliability and ROC**

Figure 10 shows the reliability diagram and ROC curves for probabilistic forecasts of 6-hourly accumulated precipitation in Mexico City for a lead time of 0–90 h and over 2017–2019. The thresholds were established based on the mean (0 < Precip < 0.47 mm) and percentile 95th (Precip > 2.4 mm). The event was considered to occur if the forecasted probability was greater than 0.5. For Precip > 0.47 mm, over-forecasting is presented for events associated with high probabilities (Forecasted Probability > 0.6). If we now turn to ROC curves (Figure 10(b)), it can be seen that the curves lie a little bit closer to the 45\(^\circ\) diagonal, so the performance of the model was not as good as for the first zone analyzed. However, the AUC values obtained are around 0.72, and, therefore, the model is still useful.

**DISCUSSION**

This study comprises the verification of the ECMWF’s Operational Prediction System in the Valley of Mexico Basin and Mexico City. The first question in this study sought to determine the correspondence between forecasts and observations in MVB. In a preliminary analysis, we decided to examine basic statistics, so HRES and EMEAN were evaluated using non-probabilistic verification methods. With respect to this first research question, in Figure 2, it was found that that both HRES and EMEAN correspond adequately with the observed data and, in general, are successful in identifying the rainy season (MJJASON). The greatest inconsistencies occur for the upper areas of the basin, where it is observed that both HRES and EMEAN over-forecast the event, although HRES does it with greater magnitude. These results are similar to those found by Li et al. (2021), who reported a tendency to over-forecast monthly rainfall. Also, there are zones where observed rainfall was not properly identified by the forecasts (e.g. January–March). The observable differences between observations and
Figure 8 | Event 20 August 2018: analysis of the values per grid cell of the HRES and ENS forecast and maps showing the accumulated rainfall results for a horizon of 90 h. P95, 95th percentile; P05, 5th percentile.
forecasts are mainly due to the uncertainty of the predictions, although there are also other possible causes of error such as forecast resolution and the occurrence of convective rains, a poor distribution of measurement stations, arithmetic average of observed data, missing data and measurement errors.

On the other hand, nonprobabilistic forecasts (HRES and EMEAN) show a growth of errors and BIAS (Figure 3) over the months of May to November (MJJASON); this indicates greater uncertainty related to extreme event forecasts. Both deterministic forecasts presented similar results; however, EMEAN performs better in comparison with HRES.

The mean of the ensembles is closer to the possible state of the atmosphere if the ensemble dispersion is small. This means that the greater the dispersion, the greater the difference that exists between the members of the ensemble, and, thus, greater uncertainty. The results in Figure 4 are consistent with those of the previous findings; ever since the forecast was made, uncertainty is greater for the upper areas of the basin, with average spread values less than 0.5 mm. This might be because the probabilistic forecast is centered on the mean in the first hours of the forecast horizon. However, this spread is not significant, so it can be established that ENS is sufficiently precise for a lead time of 24 h.

Figure 9 | Average meteogram for a 90-h horizon rain event in Mexico City, period (MJJASON) 2017–2019. The enveloping lines correspond to the percentiles of the probabilistic forecast (ENS). Plot generated with Verif Tool (Nipen et al. 2019).

Figure 10 | Reliability and ROC curves. Time step=6 h, lt=0+90 h. Mexico City.
To evaluate the reliability and discrimination ability of the probabilistic forecasts in MVB (study zone 1), a reliability diagram and ROC curves, respectively, were used. The reliability diagram in Figure 5(a) shows that ensembles over-forecast scarce precipitation events (Precip < 5 mm) are associated with low probabilities of occurrence (Forecasted Probability < 0.4). For events greater than 5 and 10 mm, the forecasts are biased with over-forecasting of events associated with high probabilities (Forecasted Probability > 0.6). One unanticipated result was the low reliability for a threshold of 20 mm, which indicates mostly over-forecasting; yet, it is possible that these results are not a true representation of the event because the shape of the curve indicates that it is probably under-sampled (e.g. Ebert 2007). Despite this, ROC curves in Figure 5(b) showed that the ensembles’ discrimination ability is significantly good, since the curves lie closer to the upper left corner of the plot. In addition, the area under the ROC curves resulted in values around 0.85, which is indicative of a good forecasting system and a satisfactory performance of the predictive model.

The second question in this research is whether forecast resolution is sufficient for flood prediction in Mexico City; moreover, it is important to analyze the behavior of the forecasts over the time horizon to establish a limit in which it remains reliable, and thus, to be able to issue an alert with adequate anticipation. In accordance with this, the study of isolated events allowed to analyze the spatial behavior of the forecasts through a lead time of 90 h (Figures 7 and 8). The most important results of this analysis are that, in general, ENS has a tendency to over-forecast light rain events, which is consistent with previous results. In addition, this analysis corroborates that HRES performs better in the first hours, showing a level of participation of up to 60% for the first 24 h of the time horizon (see Richardson et al. 2015). However, the longer the time horizon, the ensemble participation will be greater due to the uncertainty of the forecast.

According to Figure 9, HRES forecast is sufficiently accurate in the first 36 h; after that time, it begins to over-forecast events, so it is no longer reliable; it is then when ENS presents the advantages, since they do represent the probability distribution of a real event. This finding is consistent with Gouweleeuw et al. (2005), who established that some members of the ensemble may be representative of extreme events and who are within the 95th percentile of the distribution (e.g. zones 3, 5 and 10 in Figure 7; zones 4, 9 and 11 in Figure 8).

The results of the reliability diagram and ROC curves in Figure 10 showed that the EPS can predict the accumulated rainfall in 6 h and that these predictions can be useful for a forecast horizon of up to 90 h. In accordance with Figure 10(b), the performance of the predictive model was good enough since areas under the ROC curves indicate that the model is still useful to predict the event over a lead time of 90 h.

CONCLUSIONS

The purpose of the current study was to examine the potential use of ECMWF’s operational forecasting system to predict rain events that cause flooding in CDMX. The results indicate the necessary participation of a probabilistic forecast that is representative of the probability distribution of the future states of the atmosphere and, therefore, the inclusion of extreme events.

Regarding probabilistic verification, the ensembles showed a greater dispersion in the upper areas of the Valley of Mexico Basin.

The visual analysis of flood events (June 27, 2017 and August 20, 2018) showed that the ensembles are representative of the observed event in all grid cells; and despite their resolution, the ECMWF’s EPS is a very useful tool for doing a probabilistic analysis of the occurrence of an event in Mexico City.

To complement the above, the average meteogram in CDMX shows that the EPS constitutes a good representation of the possible scenarios of the atmosphere along the time horizon. As a result, the observed event is not statistically different from any other member of the ensemble. In addition, the first 36 h of the forecast are the most reliable, after that time, forecast uncertainty increases.

Finally, ENS has good reliability for thresholds of 5–20 mm of rain accumulated in 24 h and manages to adequately represent the mean and 95 percentile of rain accumulated in 6 h for a forecast horizon of 90 h. On the other hand, the logit model performed satisfactorily with areas under the ROC curves (AUC) of 0.85 and 0.72 for the MVB and CDMX, respectively.

The insights gained from this study may be of assistance to motivate the use of forecasting products for the implementation of flood warning systems at the local level, since one of their greatest advantages is the help they offer for decision-making using limited and user-friendly data. These findings provide additional evidence of the reliability and precision of ensemble forecasts at a temporal resolution of 6-hourly accumulations, which can be of assistance for flash flood analysis and which can
provide support to the city’s civil protection services, allowing them to attend to those vulnerable areas with sufficient time in advance.

The strengths of the study are forecast verification using high-resolution, real-time observed data and the evaluation of forecasts for a lead time of 90 h, which allowed establishing the most reliable lead time of forecasts for the study area.

With regard to the research methods, some limitations need to be acknowledged: the observed data were considered as true, but it must be taken into account that the different measuring devices also present a certain degree of error. Although the verification process can be replicated, having observed real-time data can be a limitation for forecast verification at lower temporal resolutions (time < 12 h) for other study zones. In addition, no forecasting system can achieve perfect results, and as indicated by Murphy (1993), the forecasting process culminates in the formulation of judgments by the forecaster or the user about the occurrence or not of an event of interest.

It is important that the choice of thresholds is established with the necessary care in order to avoid possible false alarms or unforeseen events that would inevitably have social and economic repercussions (Siccardi et al. 2005). Moreover, it is advisable to make a combined use of the HRES and ENS products, since nonprobabilistic forecasts fail to identify threats.

More broadly, research is also needed to determine the appropriate precipitation thresholds for issuing alerts, as well as the probability associated with the occurrence of an event for decision-making. A greater focus on the use of predictive modeling could produce interesting findings that help to improve the ensemble probabilities or in their case, calibrate these probabilities.

In terms of directions for future research, it will be important to explore using the EPS for smaller timescales (e.g. 1 h), which can be carried out with the combination of ensemble forecasts with a hydrodynamic model for the generation of probabilistic pluvial flood maps, or a modeling of the drainage system, which allows better risk communication and the identification of the most vulnerable areas to flooding and protection in urban zones.

ACKNOWLEDGEMENTS
This work was supported by the ECMWF and its access license (access to the ECMWF systems by means of an interactive client software that allows the licensee to retrieve archive products) granted to use archival products, directly contributing to the first author PhD research. The first author thanks the National Council of Science and Technology (CONACYT) and the National Autonomous University of Mexico for the sponsorship and support in the realization of this project. CONACYT had no such involvement in the decision to submit the article for publication. I would like to thank my supervisor Adrián Pedrozo-Acuña for their contributions and help in the development of this project. Secondly, the first author directly thanks Dr Thomas Nipen for his guidance in using Verif tool for the application of verification metrics. Most importantly, the first author deeply appreciates the help, guidance and support provided by ME Marcela Severiano, and three anonymous reviewers are thanked for their careful reviews of this manuscript.

DATA AVAILABILITY STATEMENT
Data cannot be made publicly available; readers should contact the corresponding author for details.

REFERENCES


León, C. & Neri, C. 2010 *Pobrezia Urbana Y Cambio Climatico Para La Ciudad De Mexico (Urban Poverty and Climate Change for Mexico City)*. Mexico.


Reinoso, E., Torres, M. A., Jaimes, M. Á., Franco, V., Domínguez-Mora, R. & Pozos Estrada, Ó. 2012 In: Escenarios de Pérdidas por Inundación Pluvial en la Ciudad de México Ante la Falla de Algún Componente del Sistema de Drenaje Principal (Scenarios of Losses due to Pluvial Flooding in Mexico City due to the Failure in the Main Drainage System), 1st edn. (Gómez Mora, O., ed.). Instituto de Ingeniería, Unidad de Promoción y Comunicación, México.


Secretaría de Gestión Integral de Riesgos y Protección Civil 2019 Atlas de riesgo – Inundaciones (Risk Atlas – Floods), Gob. la Ciudad México. Available from: https://datos.cdmx.gob.mx/explore/dataset/atlas-de-riesgo-inundaciones/information/?dataChart=eyJxdWVyMjVzIjpbeyJjb25maWciOnsiZGF0YXNldCI6ImF0bGFzLWRlLXJpZXNfY3J1dGl0bWVudGFjaW9uZXMiLCJyZWRkIjoidHlwZV9saWtlIn0sImNvbHVtbCI6dHJ1ZSwidHlwZSI6ImNvbHVtb (accessed 17 September 2019).


First received 29 May 2021; accepted in revised form 1 November 2021. Available online 15 November 2021