ISSN 2007-9737

A Data Machine Learning-driven Approach to Explore Resilience and Sustainability of Mexico's Water Resource Management

Adrián Landaverde-Nava, Cristian Gonzaga-López, Michael Steven Delgado-Caicedo, Evelyn Geovanna Pérez-Gómez, Jesús Yair Ramírez-Islas, Luis Gerardo Lagunes-Nájera, Deborah Tirado-Hernández, Elisabetta Crescio, Miguel Gonzalez-Mendoza*

Instituto Tecnológico de Monterrey,
Escuela de Ingeniería y Ciencias,
Mexico

{A01745052, A01745134, A01652281, A01368866, A01275404, A01275215, A01746806, elisabetta, mgonza}@tec.mx

Abstract. Nowadays, Mexico faces several water challenges. Among them, the most critical are water scarcity, water quality, and flood management. These issues are exacerbated by climate change, population growth, and insufficient infrastructure, posing significant risks to communities across the country. This project leverages cutting-edge technologies, including artificial intelligence models and spatial inference in Python, to tackle these challenges and enhance the resilience of Mexican communities. A time series model, developed using Prophet, was trained to forecast drought levels in every state of Mexico for the period 2024-2026, using available historical data. This predictive approach aims to support policymakers in preparing for and mitigating the effects of prolonged droughts. In addition, groundwater quality was analyzed using Ordinary Kriging Interpolation and clustering techniques, enabling the identification of areas most at risk of water quality deterioration. Finally, an image segmentation deep learning model was implemented to analyze images, focusing on detecting large bodies of water and mapping flooded regions. Together, these tools offer a comprehensive strategy for managing Mexico's pressing water-related challenges.

Keywords. Water scarcity, water quality, flood segmentation, artificial intelligence.

1 Introduction

Mexico is grappling with significant water-related challenges that jeopardize both its ecosystems and its population. Issues such as water scarcity, pollution, distribution inefficiencies, flood management, and climate change adaptation require urgent and innovative solutions. Approximately 30% of Mexico's population faces severe water scarcity. This affects over 60 million people who live in regions where water supply is critically insufficient [10]. Mexico ranks among the top 10 countries with the highest water stress levels, indicating severe pressure on its water resources (Aqueduct Project, 2020).

Urban areas like Mexico City experience acute water shortages, with some neighborhoods receiving water only a few times a week [4]. In addition to scarcity, water quality is a significant concern. Approximately 70% of Mexico's rivers and 80% of its groundwater are contaminated with pollutants, including heavy metals and agricultural runoff [5].

This widespread contamination poses serious health risks and impacts millions who rely on these water sources for drinking and agriculture. The financial burden of addressing water pollution, which includes treatment and health costs, surpasses 1.2 billion USD annually [9]. These figures underscore the need for advanced methods to monitor and improve water quality.

Flooding represents another major challenge, exacerbated by climate change and inadequate infrastructure. Extreme weather events frequently cause severe flooding, leading to substantial economic and human losses. Recent years have seen billions of pesos in damages and recovery costs due to flooding [3]. The recurrent nature of

these floods highlights the necessity for improved flood management and prevention strategies.

To address these pressing issues, this work explores the potential of Artificial Intelligence (AI) as a transformative tool for water management in Mexico. By leveraging AI-driven approaches for data integration, modeling, and visualization, we aim to develop effective solutions that enhance the resilience and sustainability of Mexico's water resource management. The focus is on three critical areas: water scarcity, water quality, and flood segmentation.

- **Water Scarcity:** Usage of monthly data of droughts of every state of Mexico to make a forecast model using Prophet. This tool can identify changes in trend and seasonality, which converts it into a great forecasting tool in both precision and explainability. The decomposition of this time series analysis makes it useful to see which months are more prone to droughts, and the general trend in the last years.
- **Water Quality:** Usage of subterranean water quality data of the main pollutants of several water monitoring stations in Mexico. From this data, firstly it was applied a clustering method to obtain continuous values of water quality in each station in order to be more comparable among each other. Secondly, an interpolation method to extend the values of water quality to unmeasured regions and a general outlook of the quality in a certain region.
- **Flood Segmentation:** Usage of images of floods in order to create an image segmentation model using DeepLab V3 to identify flooded areas. This tool helps to have an automated way to identify zones in which water is being accumulated, in order to have a quick response from the first moment this problem is identified.

Through data cleaning, integration, and formatting, followed by modeling and validation, culminating in visualization, the project aims to develop and demonstrate effective AI-driven solutions.

2 Related Works

Rosanna Bonasia, specialized in Computational Fluid Dynamics and risk analysis related to natural and engineering phenomena through the application of numerical models and statistical analysis, has focused on applying mathematical models to fluid-related topics, such as air and water. In the case of floods, Bonasia, studied a city in Mexico (Villahermosa Tabasco) due to its frequent flooding. These flood discharges using a 2D shallow water model to estimate the increase in water depths in the city from 1992 to the 2019 [2].

This gives a useful tool for planning for the most frequent locations of floods. To evaluate the influence of urban expansion on flood levels, three future urbanization scenarios were proposed based on the forecasted urban growth rate for 2050. The results confirmed that the change in land use in the hydrological basins is the main factor explaining the increase in flood events observed in recent years.

This study also provided useful information for the future planning of the city, which could help minimize the impact of flooding in Villahermosa. In the case of droughts, Mexico's government has its own drought monitor, in which they graph the most recent levels of droughts throughout the country.

They reported monthly droughts in each county, with detail of the amount of water and the suggested activities of people to save water during droughts. However, this is a descriptive approach, since there is no forecast of the values of droughts, or an expected amount depending on the year or the month [5]. With water quality issues there is a similar situation as described before, since Mexico's government provides lots of historical information about superficial and subterranean water quality, as well as detailed reports of the current situation of many water bodies.

However, these works are merely descriptive and not predictive [4]. In other words, there is lots of information related to the current situation of water in Mexico, however, there isn't much analysis of how this situation might affect in the future.

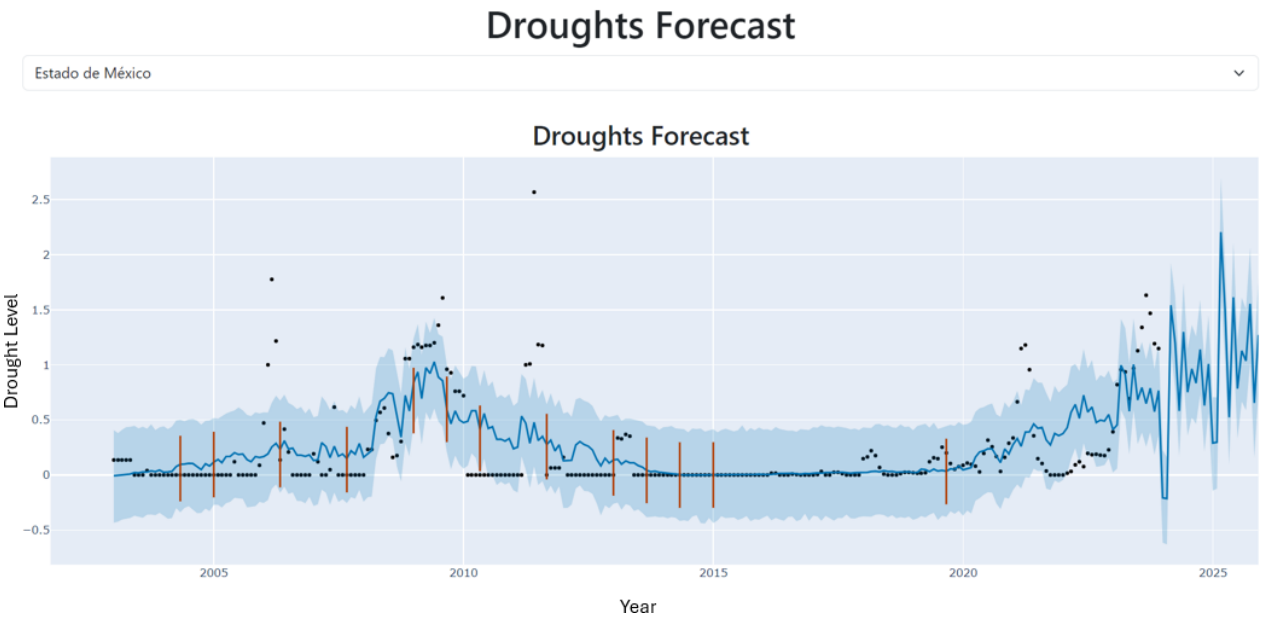

Fig. 1. Forecatsing of droughts in Estado de Mexico

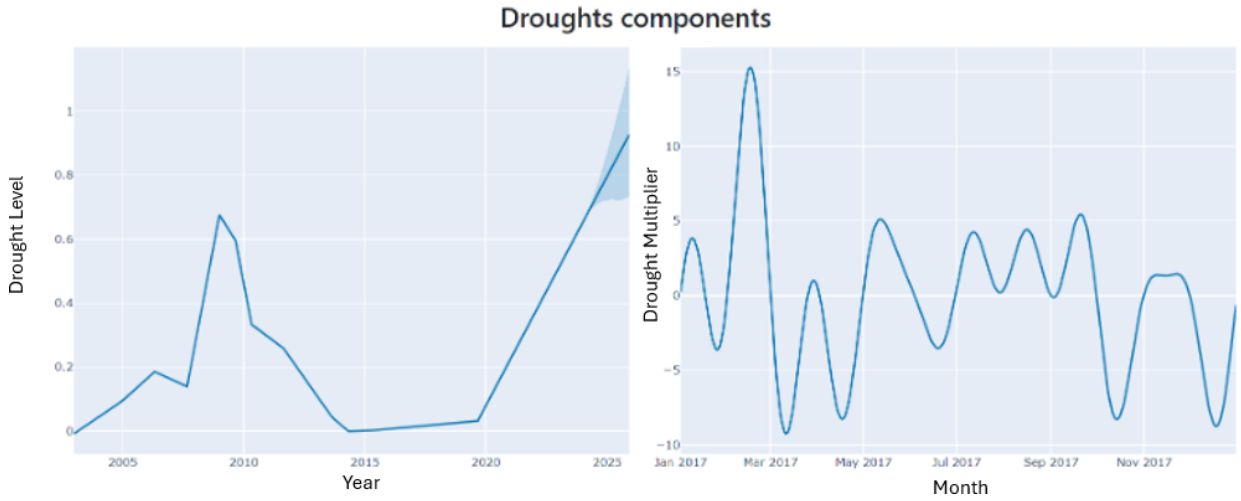

Fig. 2. Time series decomposition

3 Methodology

The present work explores the application of Machine Learning models in the management of water resources in Mexico, with a particular focus on: Drought Analysis, Flood Management, and Water Quality Monitoring. By integrating these models, the aim is to improve the resilience of Mexican communities to current and future water challenges, promoting sustainable and efficient management of water resources.

A series of steps were undertaken for the implementation of the models for the analysis of the three aforementioned approaches: droughts, floods, and water quality. The whole process was made entirely using Python with a diverse range of

libraries such as Numpy, Pandas, Plotly, Prophet, Scikit Learn, and Tensorflow.

3.1 Drought Analysis

In this section, a model trained to forecast the level of droughts in 2 years in every state of Mexico is implemented. Also, it identifies the seasonality and trends in the past. This approach is useful for 2 main reasons: On one hand, by doing the forecast of droughts, it can be suggested some kind of planning to save and keep water when it would be most needed.

On the other hand, by understanding the trend and seasonality of the data, it can be identified periods when droughts are more prone to happen. In other words, understanding the past of the droughts can be used to get ahead and plan for these intervals of water scarcity.

3.1.1 Data

A database containing the level of droughts in every county of Mexico was used. The granularity of the data consisted of biweekly observations from 2003 to 2023.

The level of droughts was measured in 5 levels, from D0 (No droughts) until D4 (Extreme droughts). The cleaning of the data consisted of grouping the data into monthly observations per state and obtaining the mean of the grouped values. Furthermore, the data was joined into a Shapefile file to get the geometries of each state for later visualization.

3.1.2 Visualization

Using Plotly, a Choropleth map with an animation of the level of droughts per month in each State of Mexico was made. By doing so, exploring how droughts have behaved all over the country for the last 20 years.

3.1.3 Modeling

An approach of time series forecasting using Prophet was used. Prophet is a library made by Meta based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. This additive behavior is shown on Equation 1, where the model forecast a value y, at a time t by summing $g(t)$ (growth component), $s(t)$ (seasonality component), $h(t)$ (holiday component), and ϵ_t (error term) [7]:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t. \tag{1}$$

In the case of the trend, which was selected as linear growth, is described in Equation 2, which resembles the basic linear equation $y = mx + b$, but with added terms that allow for changes in the trend. And the seasonality term is described in Equation 3, which consists of a Fourier series, or in other words, a summation of multiple sinusoidal curves determined by the amplitude, phase, and periods of each component curves [7]:

$$g(t) = (k + a(t)^T\delta)t + (m + a(t)^T\gamma), \tag{2}$$

$$s(t) = \sum_{n=1}^{N}\left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right)\right). \tag{3}$$

The model was trained for every state with the same parameters: multiplicative seasonality, strong seasonality effect (seasonality_prior_scale=10), and mid changes in trends (changepoint_prior_scale=1). With this approach, it was obtained an approximate MAPE value of 10% in each state.

Finally, the results included two interactive graphs. One with the forecast for the following two years, and another one with the description of the effect of the general trend and the monthly seasonality of the time series.

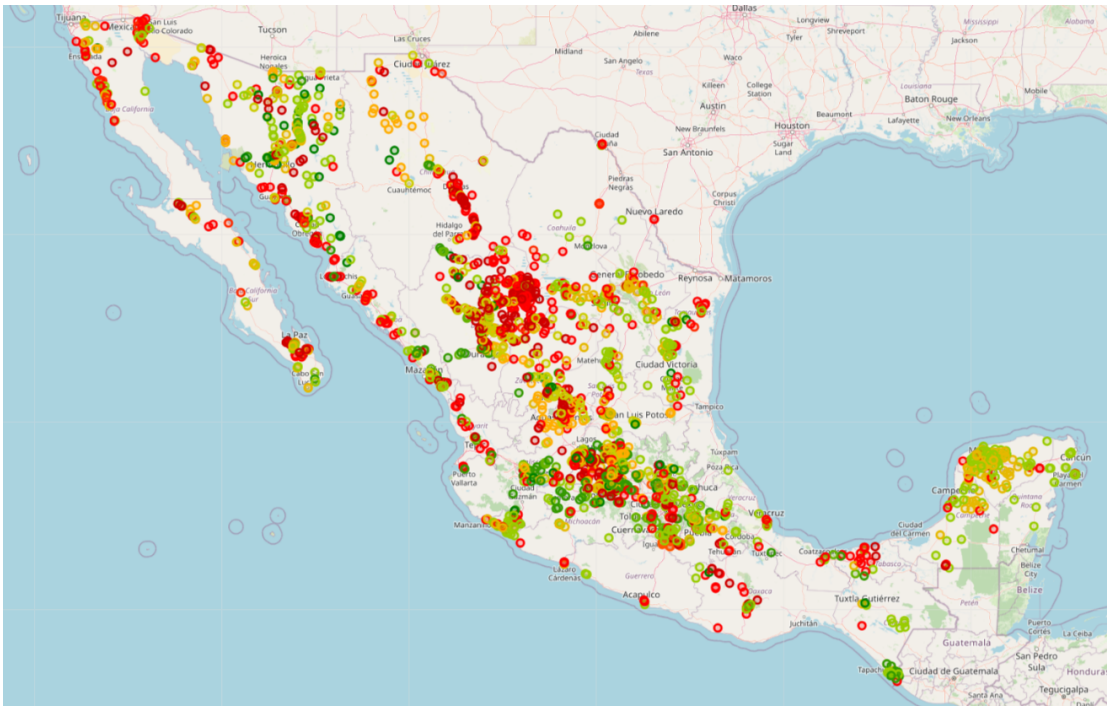

Fig. 3. Clustering of water quality per station

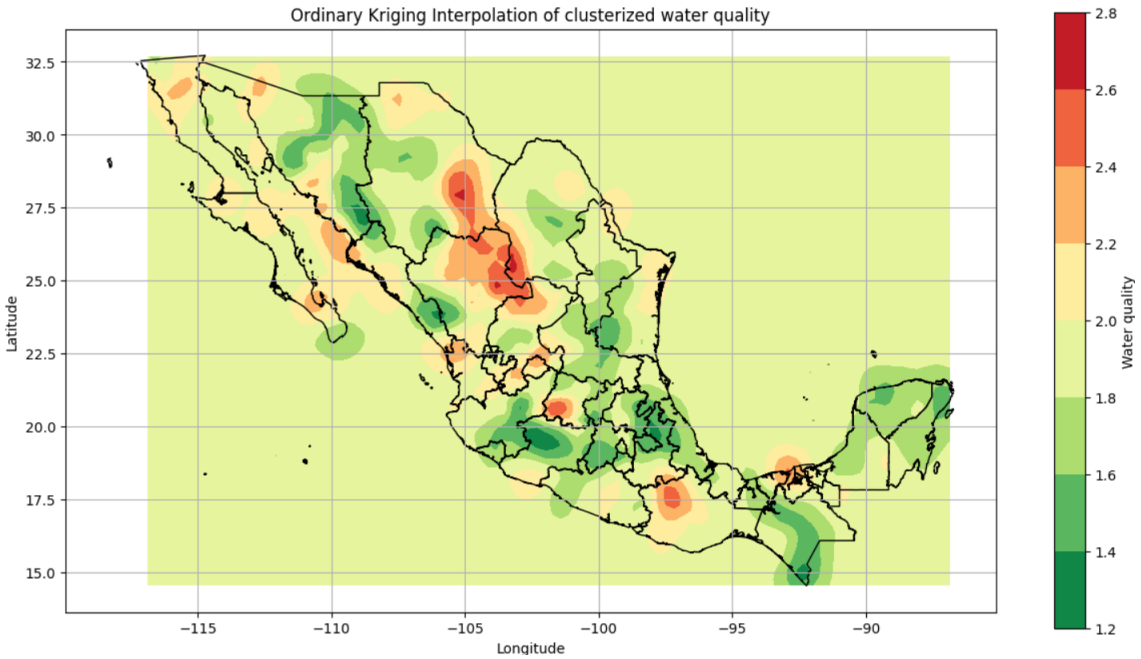

Fig. 4. Ordinary kriging interpolation of water quality from clusters

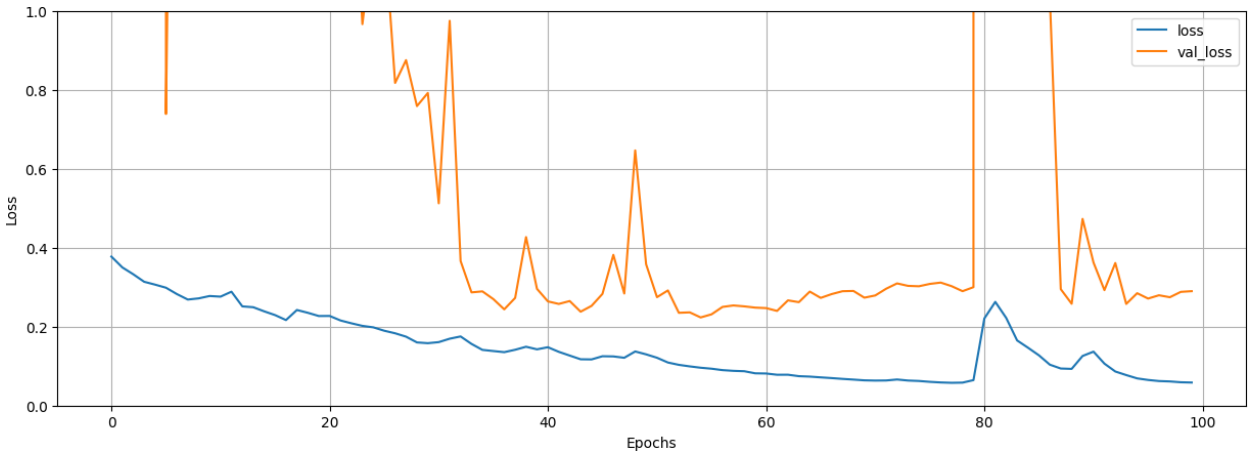

Fig. 5. `Loss` against `val_loss` of the model training

3.2 Water Quality

A clustering of water quality from multiple subterranean water monitoring stations and an interpolation of the quality from these sources into all the country were implemented. This approach would be mainly used to identify zones that have similar values in some pollutants but different classifications in order to group these places and have an idea of which places are more prone to worse its water quality. Also, these features could be extended to other regions of the country, since subterranean water is shared among multiple regions, the quality is extended to the entire region.

3.2.1 Data

The groundwater quality dataset was retrieved from the National Water Commission (CONAGUA, by its acronym in Spanish). It refers to groundwater quality indicators at the national level from 2012 to 2022. It consists of 775 observations and 56 variables, including both qualitative and quantitative data of diverse pollutants in water, as well as a general classification of water quality.

The pollutants measured in this dataset included: Total Alkalinity, Total Arsenic, Total Cadmium, Fecal Coliforms, Conductivity, Total Chromium, Total Hardness, Total Iron, Total Fluorides, Total Mercury, Total Manganese, Nitrate Nitrogen, Total Lead, Total Dissolved Solids (Agricultural Irrigation), and Total Dissolved Solids (Salinization).

Furthermore, we applied a data cleaning process mainly eliminating non-numerical characters, casting strings to numbers, filling empty values with the minimum value of the data, and normalizing the data to have the same range of values.

3.2.2 Modeling

Given the previous data, it was favorable to convert the general data quality of the data from discretized values to continuous. This would make each zone more comparable to each one, and identify zones with overlapping behavior of water quality.

Firstly, using the numeric normalized values of quality from the 14 different pollutants, a Manifold Learning algorithm was used to reduce the dimensions of the dataset.

Since a later task of the model was to cluster the data, it was selected a t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm, which is a nonlinear dimensionality reduction algorithm, described as the similarity of datapoint x_j to datapoint x_i is the conditional probability $p_{j|i}$, that x_i would pick x_j as its neighbor with an added term of perplexity Equation 4 [6].

So, this model is optimized for both dimension reduction tasks and identification of clusters. Achieving the reduction of dimensions of the

dataset to 2 dimensions and getting clear separation of the different clusters:

$$p_{j|i} = \frac{\exp\left(-\frac{||x_i - x_j||^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{||x_i - x_k||^2}{2\sigma_i^2}\right)}. \quad (4)$$

Then, a clusterization of the two dimensional data was done, using DBSCAN. This algorithm consists in assigning data points to clusters that share a neighborhood with a distance ϵ, this neighborhood is described in Equation 5 [8].

After doing so, 14 different clusters were obtained. However, after comparing the clusters, some of them had data of only one water quality level, but others had more than one level. So, the mean of the water quality per cluster was calculated, obtaining continuous values from 1 to 3:

$$N_\epsilon x = B_d(X, \epsilon) = \{y | \delta(x, y) \leq \epsilon\}. \quad (5)$$

Finally, an Ordinary Kriging Interpolation was performed in order to interpolate these continuous water quality levels to more parts of the country. This method assumes that the distance between points reflects a spatial correlation that can be used to explain variation in the surface. It weighs the surrounding measured values to derive a prediction for an unmeasured location. The general formula is formed as a weighted sum of the data, shown in Equation 6:

$$\hat{Z}(s_0) = \sum_{i=1}^{N} \lambda_i Z(s_i), \quad (6)$$

where $Z(s_i)$ is the measured value at the ith location, λ_i is an unknown weight for the measured value at the ith location, s_0 is the prediction location and N is the number of measured values [1]. This was made in order to detect water quality in zones rather than specific spots, given that subterranean water is shared between regions.

3.3 Floods

This model consisted of an image segmentation deep learning model which identifies large bodies of water in an image, specifically, flooded regions. This approach to water detection can potentially assist sectors of the population residing in flood-prone areas by promptly alerting law enforcement agents and facilitating preventive measures based on regional data to mitigate flood impacts in the area.

3.3.1 Data

The dataset consisted of 290 pair of images of floods from different regions obtained from Kaggle. Each pair of images consisted of the actual image of the region with the flooded area and a mask that identified the region of water.

3.3.2 Modeling

The implemented model is a variant of the DeepLabV3 model, designed for semantic image segmentation, specifically adapted for detecting flooded areas. The architecture of the model can be described as follows:

- **Input Layer**: The model input is an image of size 256×256 with 3 color channels (RGB).
- **Backbone**: `ResNet50` pretrained on `ImageNet` is used as the base network to extract high-level features from the images. Specifically, features from the `conv4_block6_2_relu` layer are utilized.
- **Atrous Spatial Pyramid Pooling (ASPP)**: To capture contextual information at different scales, an ASPP module is implemented, consisting of:
  - **Image Pooling**: Global `AveragePooling2D` is performed, followed by a convolution (`ConvBlock`) and `UpSampling2D` to adjust the size of the input feature.
  - **Atrous Convolutions**: Convolutions with different dilation rates (1, 6, 12, and 18) are applied to capture features at different resolutions.

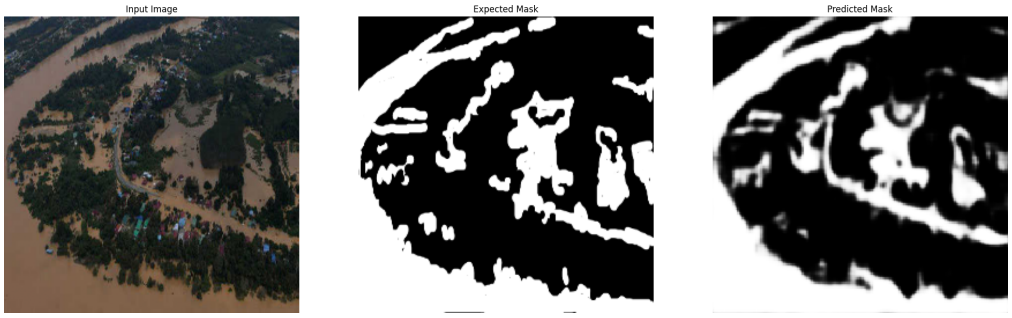

Fig. 6. Original image, original mask, predicted mask

  - **Concatenation**: Outputs from the previous layers are concatenated and passed through an additional convolution (`ConvBlock`).
- **Low-Level Features**: Additional features are taken from a lower layer of the backbone network (`conv2_block3_2_relu`) and passed through a convolution block (`ConvBlock`).
- **Feature Fusion**: Low-level features and features processed by the ASPP module are concatenated and passed through several convolutional layers (`ConvBlock`).
- **Upsampling**: The output is passed through an `UpSampling2D` layer to restore the original size of the input image.
- **Output Layer**: Finally, a convolutional layer with a `sigmoid` filter and activation generates the binary segmentation map, indicating the areas of interest (flooded areas).

The model is compiled with the `binary_crossentropy` loss function and the `Adam` optimizer. For training, callbacks such as `ModelCheckpoint` are used to save the best model and `ShowProgress` to monitor training progress. The model is trained on training images and their respective segmentation masks, with a validation set to evaluate its performance during training.

4 Results

The results obtained in the 3 types of models are shown: water scarcity, water quality, and flood management.

4.1 Droughts Analysis

As mentioned earlier, the `Prophet` model was used for time series forecasting. The trained model predicts drought levels for a 2 year period, providing a clear insight into the evolution and predictions of drought for a specific state. In Figure 1, a forecast of the droughts in Estado de Mexico for the following 2 years is shown.

The black dots are the real values, the blue line represents the forecasted values, the blue area represents the confidence interval in that period and the red lines represent changes in the trend. This forecast for the ”Estado de México” shows consistently low levels of droughts, except around 2009. However, from 2020 to the present, drought levels have increased exponentially, highlighting the current crisis that Mexico is facing.

Based on this historical behavior, the forecast for the coming years predicts unprecedented levels of drought. This situation serves as a warning of the imminent crisis related to water scarcity in multiple states of Mexico. Furthermore, in Figure 2 are shown the decomposition of the time series in Trend (left side) and seasonality (right side).

The trend portrays the general behavior of the droughts, which shows a general increase or decrease in the droughts. This plot identifies general behaviors of the data, which translates into long periods of changes in the droughts. The seasonality plot shows the multiplicative behavior of the droughts (multiplied value of the trend).

This plot is effective in identifying which months are more or less likely to have droughts. This

decomposition plot is useful to analyze the exact behavior of the droughts. Firstly, the trend analysis shows extreme changes in the general values of the droughts, representing clearly the increase in droughts around 2009, and the increase in the last years.

Secondly, the seasonality plot shows the periods in which droughts are prone to be higher, such as the peak in February, which indicated that levels of droughts are a lot higher in this month in comparison to others, showing a clear strategy to focus the efforts on reducing water consumption during this period. The forecast of every state can be seen in the following dashboard[1], made with Plotly-Dash[2].

4.2 Water Quality

After clustering the data and assigning a new water quality classification to each station, a map with continuous data regarding the quality in each station was obtained. The green points represent high values of water quality (near to 1), and the red points represent low values of water quality (near to 3). This map is shown in Figure 3.

This map show color coded continuous values of water quality on different water monitoring stations in Mexico, which portrays a more diverse explanation of the situation in these stations. For example, most of water stations of Durango have low water quality which represent a zone in which water quality in those stations and it surroundings is poor.

Other situation is seen in the North of Sonora, in which most stations have an acceptable to good quality, however there are some stations with poor values, which would indicate that these stations are outliers and should be measured again, or is a potential risk that this water might pollute surrounding water bodies, which will lower water quality in these near spots.

This potential risk is seen in other parts of the country, such as Mexico City, Estado de Mexico, Tamaulipas, and Yucatán. Posterior to realizing that most values are near similar colors, the water quality of these points could be interpolated to other zones of Mexico.

As said before, this was made using an ordinary kriging Interpolation. The results of these interpolation are shown on Figure 4. These results represent the general behavior if certain zones, which gives a general overview of the quality of grouped water bodies and how some of them might be affected by surrounding spots. The previous two maps used together are a perfect way to analyze water quality. On one hand, the interpolation map shows the overall behavior of each zone, which serves as as overview of the situation of most spots in that state. On the other hand, the clustering map shows outliers in these interpolated zones, which should be analyzed with more detail to see if this situation is an outlier or if this water might affect surrounding spots.

4.3 Floods

Finally, the model of flood image segmentation was trained, the training and validation results are shown in Figure 5. The behavior of the training and validation loss show good results, since it is avoiding both underfitting and overfitting, since both training and validation loss are lower than the initial values and have reached a plateau without increasing again. Furthermore, the efficiency of the model is portrayed in Figure 6, since the original and the predicted mask are very similar, hence, identifying accurately the zones with floods.

On the left image is shown an image of a place with a flood that covers most of the land. The image on the centre shows the original mask, which consist has manually identified these zones of water with white, and the non-water zones are shown in black. Finally, the image on the right shows the predicted mask, which has identified accurately the floods, and this is with its similarity to the original mask.

5 Conclusions

This analysis consisted in the analysis of three weak points related to water in Mexico: water Scarcity, water quality, and flood management. The importance of these three stages is that they

[1] water-ai.onrender.com/
[2] water-ai.onrender.com/

address three ways in which water might affect us: the lack of water, the quality of current water, and the excess of water. Regarding water scarcity, doing the forecast of droughts in every state of Mexico gives a general lookout of the behavior of droughts in the coming years, and as seen in the graphs, the last years show an increase in the amount of droughts, which portrays the current drought crisis, and that it shows no sign of lowering in the incoming years, at least not if the situation stays the same. Related to quality of water, the clustering and interpolation methods created a view on the specific and general outlook of water monitoring stations.

The interpolation gives an overview of the predominant quality of water in each zone, and this interpolates the quality to unmeasured zones. Furthermore, the clustered map shows continuous values of water quality in each station, which when compared to the interpolated map shows which stations are outliers. Hence, the water quality in these stations should be checked to prevent this polluted water to extend to other spots. This situation is seen in Sonora, Mexico City, Estado de Mexico, Tamaulipas, and Yucatán. Finally, the flood segmentation model provides an automated tool to identify bodies of water that indicate emerging floods in any location.

This allows for the identification of areas more prone to flooding and enables a quick response in case of a flood, thereby facilitating the implementation of precautionary measures to avoid these places or mitigate the flood. To sum up, AI linked with good data quality can give amazing results. In this analysis three areas of opportunity were identified: water scarcity, water quality, and flood management. Our solution to this work, including the code and the data, can be found in the following **GitHub repository**: github.com/AdrianLandaverde/water-ai.

Aknowledgments

The authors would like to thank the financial support from Tecnologico de Monterrey through the “Challenge-Based Research Funding Program 2022”. Project ID # E120 - EIC-GI06-B-T3-D.

References

1. **ArcGIS Pro 3.4 (2024).** How Kriging works. pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-analyst/how-kriging-works.htm.
2. **Areu-Rangel, O. S., Cea, L., Bonasia, R., Espinosa-Echavarria, V. J. (2019).** Impact of urban growth and changes in land use on river flood hazard in Villahermosa, Tabasco (Mexico). Water, Vol. 11, No. 2, pp. 304. DOI: 10.3390/w11020304.
3. **Centro Nacional de Prevención de Desastres (2025)** Monitoreo y avisos de fenómenos naturales. www.atlasnacionalderiesgos.gob.mx/.
4. **Comisión Nacional del Agua (2024).** Calidad del agua en México. www.gob.mx/conagua/articulos/calidad-del-agua.
5. **Comisión Nacional del Agua (2024).** Monitor de sequía de México. smn.conagua.gob.mx/es/climatologia/monitor-de-sequia/monitor-de-sequia-en-mexico.
6. **Erdem, K. (2020).** t-SNE explicado claramente (2020). erdem.pl/2020/04/t-sne-clearly-explained.
7. **Rafferty, G. (2023).** Forecasting time series data with prophet - second edition: Build, improve, and optimize time series forecasting models using Meta’s advanced forecasting tool. Packt Publishing.
8. **Shreiber, A. (2020).** Guía práctica del método DBSCAN. Towards Data Science. towardsdatascience.com/a-practical-guide-to-dbscan-method-d4ec5ab2bc99.
9. **World Bank Group (2020).** Global water security and sanitation partnership annual report 2020 (English). documents.worldbank.org/curated/en/969081605133747136/Global-Water-Security-and-Sanitation-Partnership-Annual-Report-2020.
10. **World Resources Institute (2024).** Data. www.wri.org/data.

ISSN 2007-9737

Article received on 28/05/2024; accepted on 24/07/2024.
**Corresponding author is Miguel Gonzale Mendoza.*