Dynamic Wind Condition Detection in Baja California, Mexico: A Machine Learning Approach for Improved Wind Management

M. Arellano-Vázquez, M. Zamora-Machado, M. Robles-Pérez, C. Minutti-Martinez, and O. A. Jaramillo

Abstract—**Accurate detection and classification of wind states is crucial to accurately assess and predict wind energy production, fire spread behavior, air quality monitoring, and understanding complex meteorological phenomena. The complex nature of wind necessitates advanced data analysis techniques to extract meaningful patterns from meteorological data sets. This study presents a wind classification and identification analysis based on a Gaussian Mixture Model (GMM) clustering method applied to a case study in "La Rumorosa", Mexico. The method analyzes four meteorological variables, relative humidity, atmospheric pressure, wind speed, and wind direction, over five years to automatically identify distinct wind conditions that can be defined as climate states, including well-known regional phenomena such as Santa Ana winds and local orographic winds. Accurate detection of wind states enables better forecasting of wind energy potential at favorable sites, wildfire risk management through predicted fire behavior and monitoring pollutant/allergen dispersal patterns. The proposed approach offers a reliable and computationally efficient method for detecting wind patterns, which extends to different geographical regions impacted by diverse wind phenomena.**

Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/9273

Index Terms—**Wind energy, Gaussian Mixture Model, Machine learning, Wind state detection, Climate modeling Renewable energy prediction.**

I. Introduction

WIND energy production has surged globally in the past decade, driving the need for data science studies to understand better and predict wind dynamics. This knowledge is critical to advance the technology behind this renewable energy source. Researchers have applied various Machine Learning (ML) techniques to wind condition analysis[1], [2], [3], focusing on clustering and classification methods to identify wind states and predict energy potential at optimal locations.

The associate editor coordinating the review of this manuscript and approving it for publication was Bruno Henrique Groenner Barbosa (*Corresponding author: Magali Arellano Vázquez*).

Magali Arellano-Vázquez, and C. Minutti-Martinez are with the Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación, Aguascalientes, México (e-mails: magali.arellano@infotec.mx, and carlos.minutti@infotec.mx).

M. Zamora-Machado is with Universidad Autónoma de Baja California, Mexicali, México (e-mail: zamora.marlene@uabc.edu.mx).

M. Robles-Pérez, and O.A. Jaramillo are with the Instituto de Energías Renovables, Universidad Nacional Autónoma de México, Temixco, México (e-mails: mrp@ier.unam.mx, and ojs@ier.unam.mx).

Several clustering methods have been employed in meteorological and energy-related studies. For example, K-means is a widely used algorithm for its simplicity and efficiency in partitioning data into clusters. However, it assumes that the clusters are spherical and equally sized, which may not align with the complex and multimodal nature of the meteorological data [4], [5]. When applied in this study, the K-means yielded results with approximately 50% accuracy compared to the empirical data, making them unsuitable for the nuanced analysis required.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) addresses these limitations by identifying arbitrarily shaped clusters and excluding noise points. This makes DBSCAN suitable for outlier detection in wind datasets but less effective for modeling overlapping clusters with high variability [6].

In contrast, the Gaussian Mixture Model (GMM) offers a probabilistic approach well-suited for datasets with overlapping clusters and complex distributions. GMM achieved more than 80% similarity with empirical data in this study, demonstrating its effectiveness. Furthermore, GMM has been successfully applied in modeling fluid dynamics, such as petroleum flows, which motivated its application to modeling wind conditions in this work [7]. By modeling the data as a mixture of multiple Gaussian distributions, the GMM accounts for the correlations between variables and captures subtler patterns that other methods may overlook [8], [9]. Since our preliminary research ([10], [11]), other recent studies have successfully applied GMM as a clustering method in the wind resource area [12], [13], obtaining better results than the previous methods used in each case.

In this study, we applied the GMM to analyze wind patterns in "La Rumorosa," located in Baja California, Mexico. This region is characterized by various meteorological phenomena, including Santa Ana winds, valley and mountain breezes, and extratropical systems. By incorporating four key meteorological variables, our approach provides a nuanced understanding of local wind conditions, offering practical insights for wind energy production, wildfire risk management, and environmental monitoring.

The wind is present in every meteorological phenomenon, transporting heat, humidity, dust, insects, and pollen. Meteorologists have classified various atmospheric circulation phenomena with varying time scales as "movement scales". "La Rumorosa" area experiences various meteorological phenomena, including mesoscale, Valley and Mountain winds, Santa Ana winds, synoptic-scale phenomena, extratropical systems, Meteorological Fronts, and influences of planetary-scale phenomena such as the Pacific High-Pressure Center and “El Niño” Southern Oscillation (ENSO) [14], [15], [16], [17].

When synoptic scale conditions are relatively calm, and no high concentration of cloudiness, wind conditions will be mainly affected and controlled by local warming, generating valley and mountain wind conditions with a well-defined diurnal cycle (Fig. 1a). In “La Rumorosa”, the time series has a directionality of day winds of approximately 280 degrees (almost due west) and night winds of 130 degrees (almost due southeast).

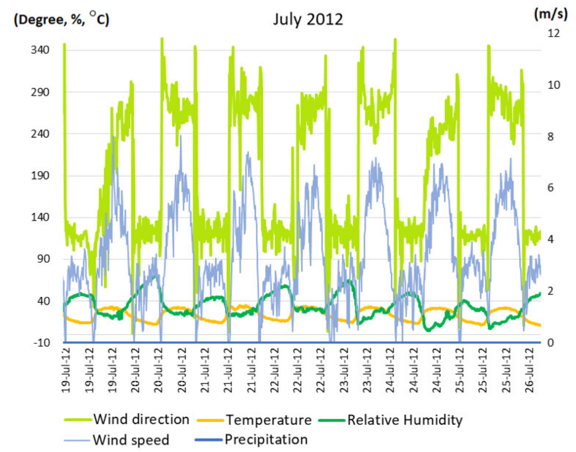
Mid-latitude cyclones generated along the polar front are present throughout the winter season. These cyclones are low-pressure systems with diameters larger than 1000 km that travel from west to east through the mid-latitudes of both hemispheres. They last from about a few days to a week.

A cyclone in the north hemisphere latitudes has a circulating pattern that moves contrary to the clock hands, with flow directed to the center (Fig. 1b). The direction of the wind oscillates between 140 and 290 degrees.

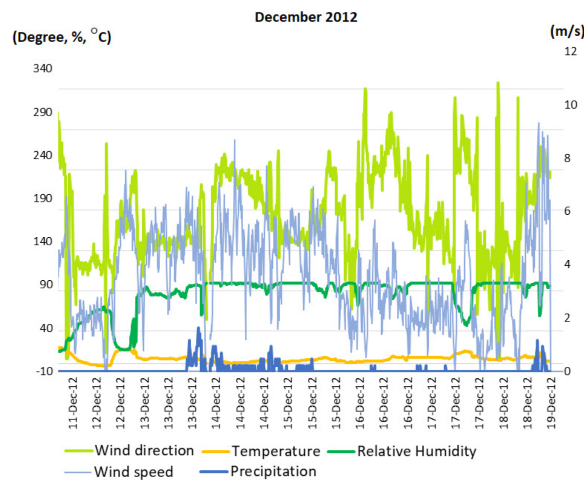
The Santa Ana Winds are among the meteorological phenomena studied most frequently in the “La Rumorosa” area. According to the literature ([18], [19], [20], [21], [22]), the Santa Ana wind (Fig. 1c) or condition can be defined as a dry wind that blows from the east or northeast toward southern California. As the air descends from the desert plain, the wind funnels through the mountains’ canyons, such as San Gabriel and San Bernardino, extending over Los Angeles, the San Fernando basin, and the Pacific Ocean. These dry winds develop in a low-pressure region formed in the Great Basin. Wind circulation flows clockwise around the anticyclonic forces that slop from the plateau. The complex topography of the land (mountains and plains) that extends from mid-south California decisively affects the characteristics that the air mass (hot and dry) acquires, influencing its direction [23].

The Santa Ana winds are limited to September to April, with December having the most significant occurrence (according to climate maps from 1968-2000). The driest months have more incidence. The frequency is less in years when the “El Niño” phenomenon is present. The Santa Ana winds have been defined in terms of speed and direction, local pressure gradients, associated temperature, and relative humidity ([24], [25], [26], [27], [28], [29], [30], [31]). The essential condition for Santa Ana winds is the existence of an anticyclone in the “Great Basin” simultaneously with a superficial low-pressure system along the coast of Baja California[27]. There are events in the same quadrant with high-speed wind conditions and low relative humidity but without increasing temperature and with low atmospheric pressure. Low-pressure systems generate these winds.

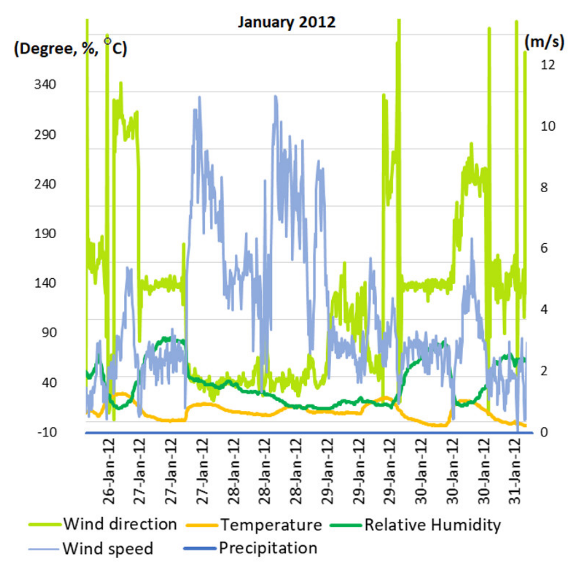
Wind data clustering in different world regions has been studied, such as the National Wind Technology Center (NWTC) in Colorado [5], where the K-means algorithm was used [4]. Four different frequent wind flows were identified, including two from the west, one from the south (which appeared during spring and summer), and one from the north



(a)



(b)



(c)

Fig. 1. Time series of local conditions of relative calm in July 2012 (a), extratropical storms in 2012 (b), and Santa Ana winds in January 2012 (c).

(mainly in summer) [5]. The method proposed in this study uses four variables of data taken on the surface (wind speed, wind direction, atmospheric pressure, and relative humidity) compared to other methods that use fewer variables [32], [33], [34]. These variables classify the Santa Ana winds and identify cold fronts in La Rumorosa [35], [27], [23], [18]. Therefore, these two types of wind states are more clearly identified with these four variables. Other variables, such as temperature and wind power [36], [11], could be added to analyze site energy generation. However, due to the availability of surface values, meteorological data on site were analyzed. It was also determined that the "El Niño" phenomenon influences the North American wind corridor, as mentioned in [37].

Recent studies have highlighted the limitations of traditional clustering methods in capturing the complex, multimodal nature of wind patterns [34], [33]. This research addresses these limitations through the implementation of a GMM-based approach that achieves over 80% accuracy in wind state classification at "La Rumorosa". By incorporating four key meteorological variables—wind speed, wind direction, atmospheric pressure, and relative humidity—the method successfully identifies both distinct wind states and previously undetected transition phases between them. The approach proves particularly valuable in distinguishing between superficially similar phenomena, such as Santa Ana winds and meteorological fronts, which share directional characteristics but differ in their pressure signatures. This comprehensive methodology demonstrates strong potential for adaptation to different geographical regions, making it relevant not only for wind energy production but also for wildfire prediction, pollutant dispersion modeling, and broader meteorological applications.

II. Methods

This study tested the GMM to identify wind patterns in each year's data. Unlike other clustering methods, such as K-Means, the GMM can consider the correlation between variables related to the phenomena being studied and use this information to identify groups within the population. In addition, the estimated parameter value has a direct interpretation with the phenomena [10]. The methodology for this study revolves around the use of *WindC1IA*, an advanced software framework developed for the automatic classification of wind states. *WindC1IA* leverages GMM to analyze multivariate meteorological data, including wind speed, wind direction, relative humidity, and atmospheric pressure. This approach ensures a robust analysis of complex and variable wind patterns.

A. Gaussian Mixture Models for Clustering

The Gaussian Mixture Models (GMMs) are probabilistic models that represent data as a mixture of Gaussian distributions, making them highly effective for clustering tasks, particularly with complex, non-linearly separable data. Each Gaussian component corresponds to a cluster, characterized by its mean (μ_k), covariance matrix (Σ_k), and a mixing coefficient (ϕ_k) that determines the proportion of data points in the cluster.

1) Key Features of GMMs: The *Number of clusters (K)* specifies the number of Gaussian components in the mixture, equivalent to the number of clusters. The *mean and covariance* define the center and shape of each cluster. The *mixing coefficients* ensure the total probability sums to 1, distributing data points among clusters.

2) Clustering with GMMs: The clustering process typically begins with parameter initialization. Subsequently, the Expectation-Maximization (EM) algorithm refines the parameters iteratively through two steps: the *Expectation Step* computes the probability that each data point belongs to each cluster (responsibilities), and the *Maximization Step,* updates the parameters to maximize the log-likelihood of the data.

3) Advantages of GMMs: GMMs offer significant flexibility, accommodating clusters of varying shapes and sizes. Unlike hard clustering methods, such as k-means, GMMs implement soft clustering, assigning probabilities to each data point to belong to different clusters. This approach enables the modeling of uncertainty, making GMMs well-suited for a wide range of data structures.

B. Model Selection Criteria

One challenge in applying clustering algorithms is the lack of a straightforward method to determine the optimal number of clusters. The Bayesian Information Criterion (BIC) is a goodness-of-fit measure in statistical modeling, often used as a criterion for selecting models from a finite set. It is based on the logarithmic likelihood function (LLF) and is closely related to the Akaike Information Criterion (AIC) [38]. When comparing these evaluations, it is possible to determine the best number of clusters. AIC provides a relative quality measure of a statistical model for a given dataset, offering a trade-off between the model's goodness of fit and its complexity. It is based on information entropy, estimating the lost information when a particular model represents the data generation process. Both BIC and AIC address the issue of overparameterization by introducing a penalty term for the number of model parameters, with BIC imposing a larger penalty than AIC [39].

C. Wind Identification through On-Site Surface Analysis

Data from automatic weather stations operated by the National Water Commission (CONAGUA), which capture data every 10 minutes, were utilized. Specifically, data from a meteorological station in a mountainous area "La Rumorosa" were studied at 1,262 meters above sea level (*m.a.s.l.*). The methodology proposed in [23] was applied to identify the Santa Ana wind condition at different meteorological stations in the state of Baja California to understand its occurrence behavior [40]. These winds exhibit asymmetric probability distributions, indicating that from late autumn to early winter, the presence of the Santa Ana condition is more likely [27]. Each event's intensity, duration, and extension can vary significantly [41]. According to [35], the wind in northern Baja California can create jet streams that reach speeds of up to 50 *m/s* and relative humidity below 10%. For "La Rumorosa", the fastest gust (in 10 minutes) was detected at 25 *m/s*. Throughout the evaluation period from 2005 to 2014, January had the most significant number of events, followed by November. From June to August, the Santa Ana wind condition was absent. Within the 5-year evaluation period, 2011 was the year with the highest presence of Santa Ana wind [23].

D. Variable Selection and Data Analysis

The selection of wind speed, wind direction, relative humidity, and atmospheric pressure as key variables for this analysis is based on their fundamental role in characterizing complex atmospheric phenomena. Together, these variables provide a comprehensive representation of wind states, capturing both the mechanical aspects of air movement and the thermodynamic conditions that drive wind patterns. The interrelation between these variables is particularly crucial for identifying distinct wind phenomena; for example, the combination of wind direction and atmospheric pressure helps distinguish between Santa Ana winds and meteorological fronts that might otherwise appear similar when examining wind characteristics alone. This multivariate approach enables a more nuanced understanding of the complex atmospheric dynamics present in the study region.

E. Theory and Calculations

For analyzing the data obtained from meteorological stations, the Python function *sklearn.mixture.GMM* was used, which includes a GMM representation of the probability distribution based on the maximum likelihood principle. We considered measurement data from 2010 to 2014, in addition to empirical analysis of wind, Santa Ana winds characterization, and Meteorological Fronts in the Santa Ana region. An observational analysis, using an annual time series for five years, was performed for empirical diagnosis. The Santa Ana wind classification parameters were a wind speed of 4 *m/s*, sustained wind speed in the northeast quadrant, decreased relative humidity, and increased temperature. The proposed method detected these wind states using local atmospheric pressure.

The increase in wind speed and the decrease in relative humidity from values near 100% to those close to 10% are considered Santa Ana winds. The direction of the wind is sustained, and the temperature of the diurnal cycle becomes warmer during the night.

F. Software Implementation and Application

The analysis was conducted using *WindC1IA*, an open-source software framework developed for the classification of wind states using GMMs. The software is designed to handle large-scale meteorological datasets, providing robust clustering and visualization tools. Its main features include:

- Data Preprocessing: Automated cleaning, normalization, and handling of missing data to ensure consistency across variables such as wind speed, direction, atmospheric pressure, and relative humidity.
- Clustering and classification: Integration of GMMs for multivariate clustering, allowing for the identification of overlapping and complex wind states.
- Output and Visualization: Generation of graphical outputs and summaries of the clustering results, facilitating the interpretation of meteorological patterns.

In this study, *WindC1IA* [42] was applied to a data set with 5 years of records from "La Rumorosa". The software identified six distinct wind states, validated by empirical observations. By integrating multiple meteorological variables, *WindC1IA* provided a comprehensive understanding of the wind dynamics in the region, demonstrating its reliability and versatility for applications in wind energy and environmental monitoring.

III. RESULTS

The analysis used meteorological data from "La Rumorosa" over five years (2010-–2014). Four key variables, wind speed, wind direction, relative humidity, and atmospheric pressure, were used as input for the GMM. The two best adjustments for each year were tested, and the one that obtained the most similar state to the Santa Ana winds was chosen. This result is consistent with the six-group cluster obtained using the BIC and AIC criteria.

Before estimating the percentages of similarities between the calculated and empirical data, it is necessary to screen the data set. According to the Santa Ana wind's duration averages, the bursts should last at least three hours to consider its arrival. Therefore, all lower gusts are filtered. Using the file with dates and presences of the identified winds throughout the four years, when comparing the GMM results, the percentage of similarity between the Santa Ana winds and one of the states obtained by the GMM is shown in Table I. The data on the presence of Santa Ana winds and the calculated states by GMM are comprised of binary information only where the phenomenon (or group) is present or when there is no occurrence (zeros and ones). Calculate the similarity percentage of the empirical reference data against the one calculated by GMM clustering.

To validate our model, we developed two key metrics comparing the GMM classifications with empirical data. For each GMM group, we calculated c_i (where $i = 1, \ldots, n$, and n is the number of groups that fit the best for a given year). The first metric, the percentage of total similarity (pst), is calculated as $pst = \frac{c_i}{n_{st}}$, where n_{st} represents the total number of elements in the empirical data set. Although pst measures overall similarity, we needed a second metric specifically for Santa Ana wind detection accuracy. We developed the prediction success percentage (ppv), calculated as $ppv = \frac{u_i}{u_{em}}$, where u_i is the number of correct Santa Ana wind identifications by the GMM model and u_{em} is the total number of Santa Ana wind occurrences in the empirical data. Tables I and II present these metrics under different model configurations. Table I shows the results using only wind speed and direction, while Table II shows the results using all four variables (wind speed, direction, relative humidity and atmospheric pressure). The four-variable configuration consistently demonstrated superior performance.

Using the GMM method and the data from 2011 and six Gaussian components resulted in the clustering shown in Fig. 2.

TABLE I
COINCIDENCES OF EMPIRICAL DATA PERCENTAGE COMPARED WITH THE STATE OBTAINED USING GMM WITH SPEED AND DIRECTION

Year	pst (%)	ppv (%)
2010	89.87	88.06
2011	87.09	91.46
2012	89.21	90.42
2013	86.81	94.86
2014	88.08	87.23

TABLE II
COINCIDENCES OF EMPIRICAL DATA PERCENTAGE COMPARED WITH THE STATE OBTAINED USING GMM WITH ALL VARIABLES

Year	pst (%)	ppv (%)
2010	96.66	90.52
2011	94.25	81.55
2012	97.08	82.48
2013	95.74	88.64
2014	96.30	83.61

Fig. 3 shows the qualitative similarity between the calculated clustering of the GMM method and the empirical data. The green line illustrates the coincidences between both time series; although the GMM presents false positives, most of the data points coincide with the empirical data. As shown in Table I, 91. 46% of the presence of the Santa Ana wind was accurately predicted.

The year 2011 yielded the best results compared to the others, and even at first glance, the graphic in Fig. 3a reveals a notable similarity in the three series of data. However, in the graphic of Fig. 3b, it is possible to distinguish where the differences lie. The daily average graphic errors are shown in Fig. 3c to highlight these differences. It can be seen that, for this year, the errors were generally less than 10%.

IV. DISCUSSION

The results of this study demonstrate the effectiveness of *WindC1IA* in classifying wind states and its advantages over traditional clustering methods. The analysis used meteorological data from "La Rumorosa" over five years (2010–2014). Four key variables—wind speed, wind direction, relative humidity, and atmospheric pressure—were used as input for the GMM. This research advances the field in three key aspects: (1) improved classification accuracy through GMM implementation, achieving over 80% accuracy compared to traditional methods, (2) successful identification of transitional states between wind phenomena, and (3) reliable differentiation between superficially similar wind patterns through multi-variable analysis, particularly in distinguishing Santa Ana winds from meteorological fronts.

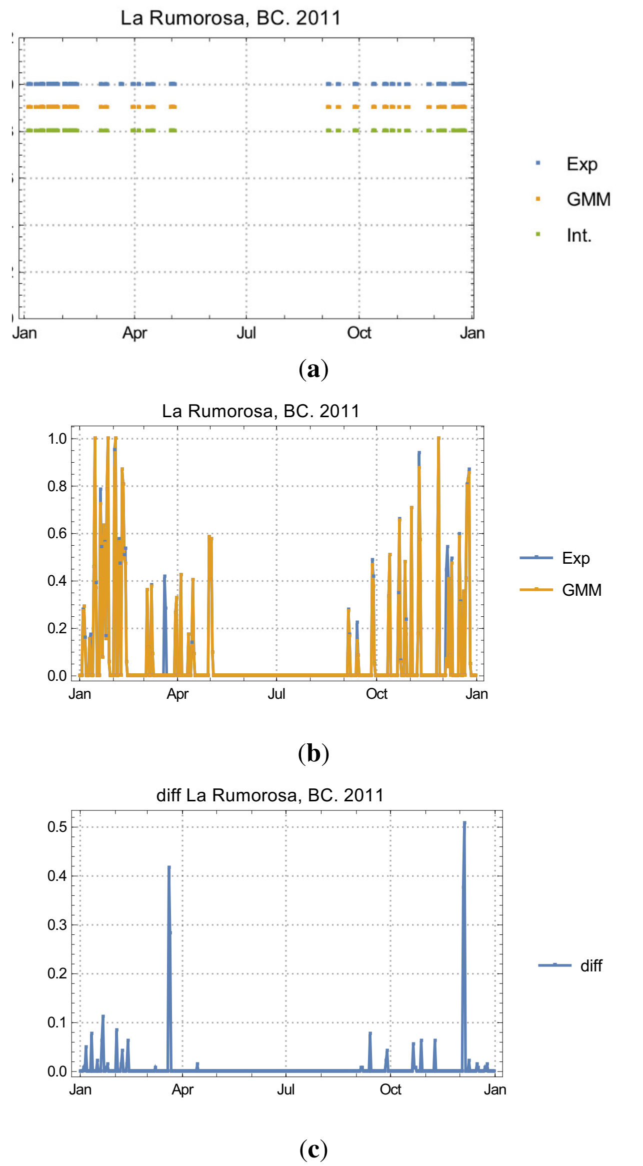


Fig. 3. (a) GMM clustering results for 2011. (b) Daily average of Santa Ana wind presence during 2011. (c) Errors in the daily average of Santa Ana winds during 2011.

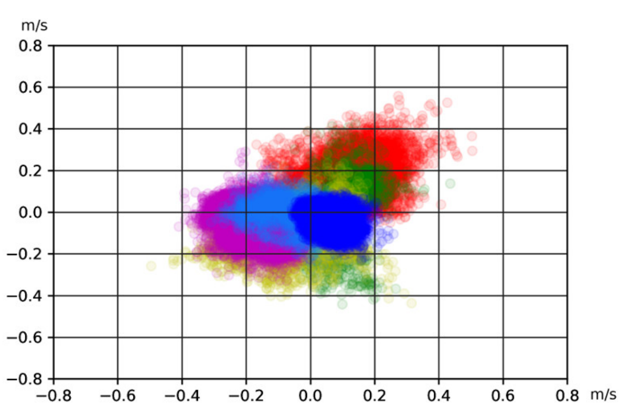


Fig. 2. clustering based on GMM with six groups, considering all data points from 2011.

A. Comparison with Other Methods

To validate the suitability of GMM, its performance was compared with K-means, a commonly used clustering algorithm. K-means achieved a classification accuracy of approximately 50% when validated against empirical observations. This limitation can be attributed to the algorithm's assumption of spherical clusters, which do not align with the region's complex and overlapping distributions of meteorological data. In contrast, GMM achieved an accuracy that exceeded 80%, confirming its ability to handle multimodal distributions and variable correlations effectively.

B. Wind State Classification

Using GMM, six distinct wind states were identified, each representing unique meteorological phenomena in "La Rumorosa". These states are as follows:

1) **Valley Winds:** Associated with high-pressure systems at night, characterized by cool descending air from the mountains and clear skies.
2) **Mountain Winds:** A low-pressure phenomenon occurs during the day, marked by an increase in warm air from the valley and an increase in humidity.
3) **Wind States Transition Clustering:** Representing the shift between valley and mountain breezes, which typically occurs during sunrise and sunset with variable wind direction and speed.
4) **Santa Ana Winds:** Characterized by strong and dry gusts originating from the northeast, often associated with high fire risk and dry conditions.
5) **Meteorological Fronts:** Wind patterns influenced by large-scale meteorological systems, with moderate speeds and consistent directional flow from the southwest.
6) **Calm Conditions:** Minimal wind activity, stagnant air, and high atmospheric stability are often linked to air quality issues.

Each state was validated using empirical observations, ensuring that the clusters correspond to physically meaningful meteorological conditions.

C. Application to Wind Energy Production

The results highlight the practical implications of *WindC1IA* for wind energy production. Valley winds, Santa Ana winds, and Meteorological Fronts were identified as optimal for energy generation due to their high wind speeds and favorable directional consistency. These findings provide actionable information for wind farm operators, enabling more efficient scheduling and use of wind resources.

D. Implications for Wind Energy and Environmental Monitoring

Identifying wind states provides operational benefits for wind energy production and actionable insights for wildfire risk management and air quality monitoring. For example, Santa Ana winds are associated with a high risk of fire due to their dry and gusty nature, while Calm Conditions indicate stagnant conditions that could exacerbate air pollution.

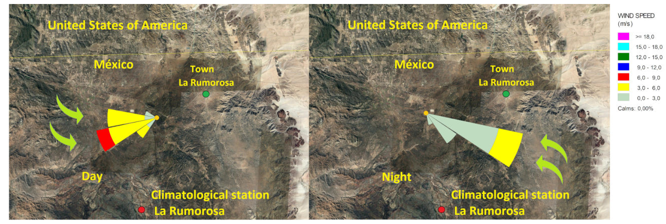


Fig. 4. Wind flow in "La Rumorosa": daytime cycle (left) and nighttime cycle (right).

E. Wind Pattern Characterization and Analysis

The detected wind conditions are similar to those shown in Fig. 1, suggesting that there should be at least five weather forecasts plus one in transition, resulting in a representative minimum of six wind states for the site. By analyzing the data in Fig. 2, four meteorological phenomena in "La Rumorosa" are identified through time series, each with its physical behavior shown below.

1) Valley and Mountain Winds: "La Rumorosa" station is located in a valley at 1,262 *m.a.s.l.*, so the diurnal wind circulation has a defined cycle. During the day, the wind is stronger, and its dominant direction is the west quadrant. Wind circulation is weaker at night, and its direction inverts to the southeast quadrant, clearly defining a valley and a mountain breeze. In Fig. 4, the wind circulation in "La Rumorosa" is observed using a compass card: daytime circulation to the left and nocturnal circulation to the right. The day is represented by the following wind condition, as observed in Fig. 5, where there is a greater frequency of occurrence in this state as local mountain and valley breeze conditions occur throughout the year. When larger-scale conditions, such as Santa Ana winds, weather fronts, or storms, are present, this state disappears until the larger-scale phenomenon ends.

At night, it is a state with more incidence, as it is well defined in the southwest quadrant, making it clear to observe in the compass card, the diurnal time series, and the graph of Fig. 6.

2) Santa Ana winds: In Fig. 7, this type of wind behavior is characterized by a sustained wind direction at less than 90 degrees, a high wind magnitude, and a decrease in relative humidity. The increase and the diurnal pattern are lost for temperature, keeping higher values at night.

Winds in the northeast quadrant are shown in Fig. 8, characterized by a sustained wind direction of less than 100 degrees, high wind speed, and low relative humidity. The temperature during the day is influenced by the temperature contained in the winds.

3) Meteorological Fronts: A state is detected at the "La Rumorosa" station, characterized by strong winds with directions towards the northeast quadrant and low relative humidity, which could be confused with the Santa Ana winds. If these variables were only considered for analysis, the sole difference of this weather forecast would be that it is formed by storms in mid-latitudes moving through the United States.

Wind direction is sustained in the northeast quadrant, with low relative humidity and high wind speed. The low-pressure center is located hundreds of kilometers from Baja, California,

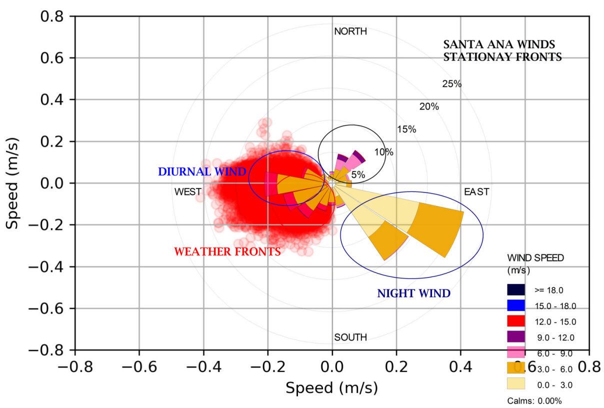


Fig. 5. Wind conditions during the day, with the highest occurrence frequency throughout the year.

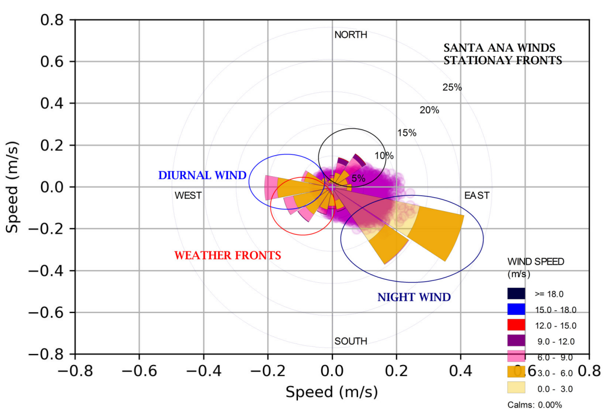


Fig. 6. Wind conditions at night, characterized by a southwestward flow.

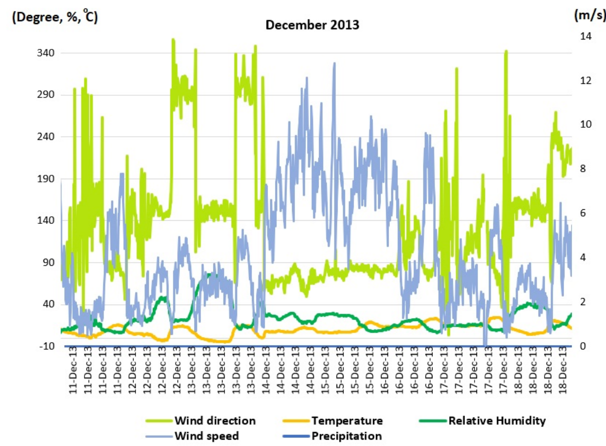


Fig. 7. Time series of Santa Ana wind behavior in December 2013.

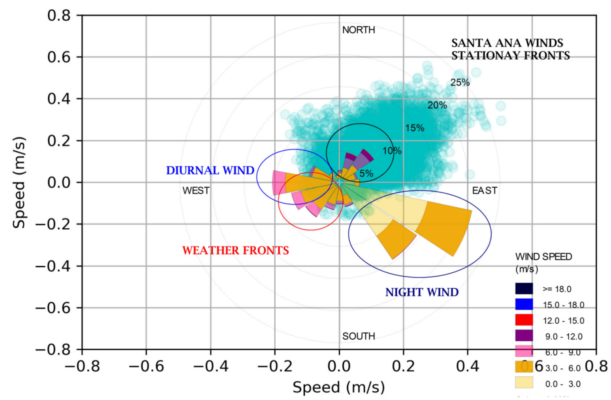


Fig. 8. Santa Ana winds impacting "La Rumorosa," with sustained direction below 100 degrees.

but due to its topography, it reaches high speeds when it affects the site. This weather state is not as frequent as others and is present from September to May.

Despite performing with two and four meteorological variables in the wind state clustering, the clustering is very similar. The only notable difference is the atmospheric pressure because when a two-variable clustering (speed and wind direction) is performed in the northeast quadrant, a very defined state is observed, which could be taken as the Santa Ana winds (Fig. 9). On the other hand, with the four meteorological variables (speed, wind direction, relative humidity, and atmospheric pressure), the previously defined quadrant in Fig. 10 separates and adds one more state that represents storms or frontal systems of low atmospheric pressure, contrary to Santa Ana winds, which have high atmospheric pressure.

4) Wind States Transition Clustering: The following states are considered to be in transition from one identified meteorological phenomenon state to another. Fig. 11a shows the day and night transition cycles since it appears they were between both states. Fig. 11b shows the transition between the night wind and Santa Ana and the meteorological fronts. Throughout winter in "El Niño" years, cyclones intensify, and Santa Ana winds decrease [27]. The following comparison is made in Fig. 11c, which shows 2010 starting with "El Niño" in winter, and Fig. 11d is for 2013, a neutral year.

V. Conclusions

The GMM method, when applied to four meteorological variables (wind speed, wind direction, relative humidity, and atmospheric pressure), enables precise detection of wind states, including differentiation between high- and low-pressure systems generated by various wind phenomena. This study successfully identified different wind conditions in "La Rumorosa", consistent with well-known meteorological events such as Valley and Mountain Winds, Santa Ana winds, and Meteorological Fronts.

The importance of this work extends beyond the wind energy sector, as accurate wind state detection is crucial for various applications, including the following:

- Wildfire Management: Identifying wind patterns and changes is essential for predicting the spread and behavior of wildfires, enabling better resource allocation and evacuation planning.

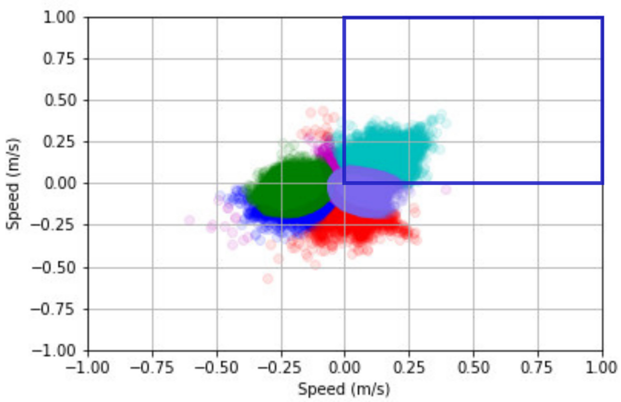


Fig. 9. Northeast quadrant of the clustering using two variables for 2010.

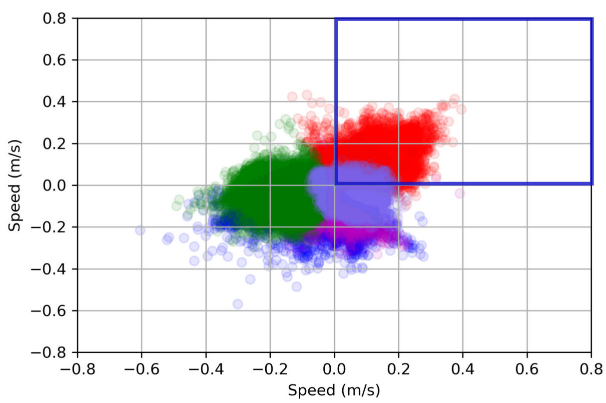


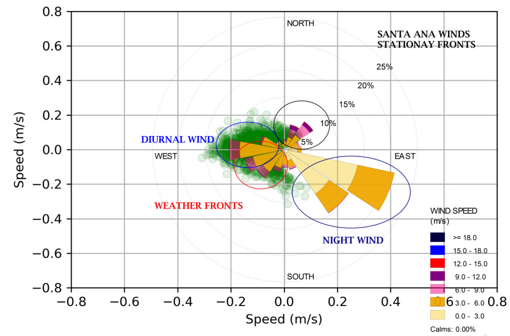
Fig. 10. Northeast quadrant of the clustering using four variables for 2010.

- Air Quality Monitoring: Wind plays a significant role in the dispersal of pollutants and allergens, and characterizing wind states can help forecast air quality and inform public health measures.
- Aviation Safety: Detecting and monitoring wind shear, turbulence, and other wind-related phenomena is vital for ensuring the safety of aircraft operations.
- Meteorological Research: Comprehensive wind data analysis contributes to a deeper understanding of complex atmospheric processes and can inform climate modeling efforts.

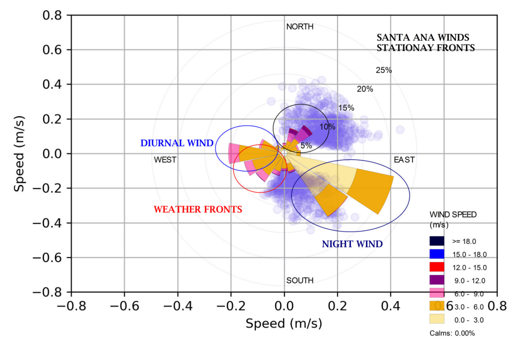
This study highlights the potential of machine learning techniques for improving wind monitoring capabilities by demonstrating the automatic detection of wind states consistent with observed meteorological phenomena. Future research aims to apply this method to more extensive data sets in diverse geographical regions, further advancing our understanding and management of wind-related processes.

ACKNOWLEDGMENTS

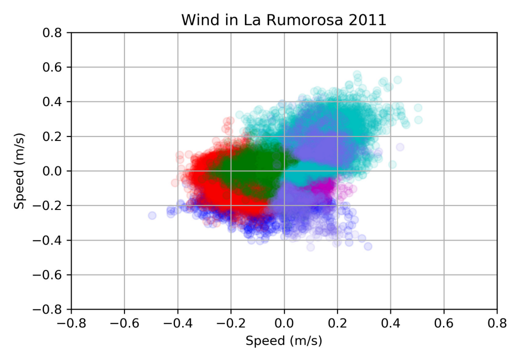
The authors acknowledge the use of Writefull and Grammarly to improve English grammar and spelling in this article. These AI systems were used exclusively to improve the clarity and



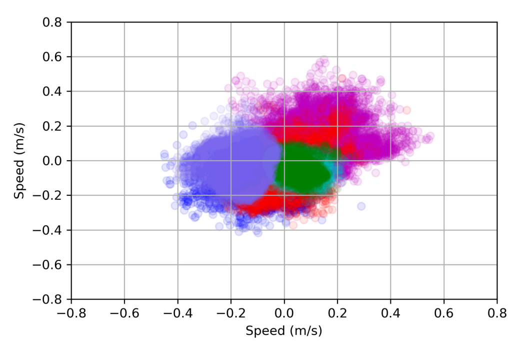
(a)



(b)



(c)



(d)

Fig. 11. (a) Shows the transition between day and night, (b) illustrates the transition between night and Santa Ana winds and Meteorological Fronts, (c) presents wind conditions in winter 2010, with the onset of the “El Niño” phenomenon and (d) depicts wind conditions in 2013, without the influence of ’El Niño.’

correctness of the text in various sections of the manuscript. The authors determined all the content and final revisions.

REFERENCES

[1] S. Buhan, Y. Ozkazanc, and I. Çadirci, "Wind pattern recognition and reference wind mast data correlations with nwp for improved wind-electric power forecasts," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 3, pp. 991–1004, June 2016. [Online]. Available: https://doi.org/10.1109/TII.2016.25430046

[2] J. Liu, M. Zhu, Z. Lin, and S. Chen, "A dynamic clustering model of wind farm based on the operation data," *Journal of Renewable and Sustainable Energy*, vol. 7, no. 4, p. 043111, 2015. [Online]. Available: https://doi.org/10.1063/1.4926906

[3] D. Alves, F. Mendonça, S. S. Mostafa, and F. Morgado-Dias, "The potential of machine learning for wind speed and direction short-term forecasting: A systematic review," *Computers*, vol. 12, no. 10, 2023. [Online]. Available: https://www.mdpi.com/2073-431X/12/10/206

[4] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2009.09.011

[5] A. Clifton and J. K. Lundquist, "Data clustering reveals climate impacts on local wind phenomena," *Journal of Applied Meteorology and Climatology*, vol. 51, no. 8, pp. 1547–1557, 2012.

[6] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 226–231.

[7] R. Najafi-Silab, A. Soleymanzadeh, P. Kolah-kaj, and S. Kord, "Electrical rock typing using gaussian mixture model to determine cementation factor," *Journal of Petroleum Exploration and Production Technology*, vol. 13, pp. 1329–1344, 2023. [Online]. Available: https://doi.org/10.1007/s13202-023-01612-7

[8] S. Konishi and G. Kitagawa, *Information Criteria and Statistical Modeling*. Springer, 2007.

[9] G. Claeskens and N. L. Hjort, "Akaike's information criterion," in *Model Selection and Model Averaging:*. Cambridge: Cambridge University Press, 2001, pp. 22–69.

[10] M. Arellano-Vázquez, C. Minutti-Martinez, and M. Zamora-Machado, "Automated characterization and prediction of wind conditions using gaussian mixtures," vol. 12468, pp. 158–168, 2020. [Online]. Available: 10.1007/978-3-030-60884-2_12

[11] A.-V. Magali, Z.-M. Marlene, M. Robles, and O. A. Jaramillo, "Favorable wind states in wind energy production at la rumorosa i wind farm," *Journal of Physics: Conference Series*, vol. 1618, no. 6, p. 062071, sep 2020. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/1618/6/062071

[12] R. Marcille, M. Thiébaut, P. Tandeo, and J.-F. Filipot, "Gaussian mixture models for the optimal sparse sampling of offshore wind resource," *Wind Energy Science*, vol. 8, no. 5, pp. 771–786, 2023. [Online]. Available: https://wes.copernicus.org/articles/8/771/2023/

[13] W. Huang, X. Zhu, H. Xia, and K. Wu, "Offshore wind energy assessment with a clustering approach to mixture model parameter estimation," *Journal of Marine Science and Engineering*, vol. 11, no. 11, 2023. [Online]. Available: https://www.mdpi.com/2077-1312/11/11/2060

[14] T.-G. R. Reyes-Coca S., *El Niño Oscilación del Sur" y los fenómenos hidrometeorológicos en Baja California: el evento de 1997/98*, 2001.

[15] E. J. Lutgens F. K., Tarbuck, *Atmosphere, The: An Introduction to Meteorology, 12th Edition*. Prentice Hall, 2013.

[16] F. Cannon, L. M. Carvalho, C. Jones, T. Hall, D. Gomberg, J. Dumas, and M. Jackson, "Wrf simulation of downslope wind events in coastal santa barbara county," *Atmospheric Research*, vol. 191, pp. 57–73, 2017. [Online]. Available: https://doi.org/10.1016/j.atmosres.2017.03.010

[17] L. M. Sheridan, R. Krishnamurthy, G. García Medina, B. J. Gaudet, W. I. Gustafson Jr., A. M. Mahon, W. J. Shaw, R. K. Newsom, M. Pekour, and Z. Yang, "Offshore reanalysis wind speed assessment across the wind turbine rotor layer off the united states pacific coast," *Wind Energy Science*, vol. 7, no. 5, pp. 2059–2084, 2022. [Online]. Available: https://wes.copernicus.org/articles/7/2059/2022/

[18] M. Zamora-Machado, A. A. Lambert-Arista, O. R. García-Cueto, O. A. Jaramillo-Salgado, E. Leyva-Sánchez, R. C. Reynaga-López, J. R. Herrera-García, N. Lizárraga-Osuna, and R. C. Anguiano-Cota, "Classification of santa ana winds for the evaluation of their wind potential in la rumorosa, baja california, mexico," *Ciencias Marinas*, vol. 49, Aug. 2023. [Online]. Available: https://www.cienciasmarinas.com.mx/index.php/cmarinas/article/view/3358

[19] J. Guzman-Morales, A. Gershunov, J. Theiss, H. Li, and D. Cayan, "Santa ana winds of southern california: Their climatology, extremes, and behavior spanning six and a half decades," *Geophysical Research Letters*, vol. 43, no. 6, pp. 2827–2834, 2016.

[20] A. Langford, R. Pierce, and P. Schultz, "Stratospheric intrusions, the santa ana winds, and wildland fires in southern california," *Geophysical Research Letters*, vol. 42, no. 14, pp. 6091–6097, 2015.

[21] M. Raphael and J. Finley, "The relationship between el niño and the duration and frequency of the santa ana winds of southern california," *The Professional Geographer*, vol. 59, no. 2, pp. 184–192, 2007.

[22] T. Rolinski, S. B. Capps, and W. Zhuang, "Santa ana winds: A descriptive climatology," *Weather and Forecasting*, vol. 34, no. 2, pp. 257–275, 2019.

[23] M. Z. Machado, "Variabilidad del potencial eólico y su relación con fenómenos meteorológicos en baja california," Ph.D. dissertation, Universidad Autónoma de Baja California, 1 2016.

[24] S. Conil and A. Hall, "Local regimes of atmospheric variability: A case study of southern california," *Journal of Climate*, vol. 19, no. 17, pp. 4308–4325, 2006. [Online]. Available: https://doi.org/10.1175/JCLI3837.1

[25] R. Castro, A. Mascarenhas, A. Martínez-Díaz-de León, R. Durazo, and E. Gil-Silva, "Spatial influence and oceanic thermal response to santa ana events along the baja california peninsula," *Atmósfera*, vol. 19, no. 3, pp. 195–211, 2006. [Online]. Available: https://www.redalyc.org/articulo.oa?id=56519304

[26] A. Bytnerowicz, D. Cayan, P. Riggan, S. Schilling, P. Dawson, M. Tyree, L. Wolden, R. Tissell, and H. Preisler, "Analysis of the effects of combustion emissions and santa ana winds on ambient ozone during the october 2007 southern california wildfires," *Atmospheric Environment*, vol. 44, no. 5, pp. 678–687, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.atmosenv.2009.11.014

[27] M. N. Raphael, "The Santa Ana Winds of California," *Earth Interactions*, vol. 7, no. 8, pp. 1–13, 2003. [Online]. Available: 10.1175/1087-3562(2003)007⟨0001:TSAWOC⟩2.0.CO;2

[28] A. Félix-Bermúdez, F. Delgadillo-Hinojosa, M. A. Huerta-Diaz, V. Camacho-Ibar, and E. V. Torres-Delgado, "Atmospheric inputs of iron and manganese to coastal waters of the southern california current system: Seasonality, santa ana winds, and biogeochemical implications," *Journal of Geophysical Research: Oceans*, vol. 122, no. 11, pp. 9230–9254, 2017. [Online]. Available: https://doi.org/10.1002/2017JC013224

[29] Y. Jin, J. T. Randerson, N. Faivre, S. Capps, A. Hall, and M. L. Goulden, "Contrasting controls on wildland fires in southern california during periods with and without santa ana winds," *Journal of Geophysical Research: Biogeosciences*, vol. 119, no. 3, pp. 432–450, 2014. [Online]. Available: https://doi.org/10.1002/2013JG002541

[30] H. Hu and W. Timothy Liu, "Oceanic thermal and biological responses to santa ana winds," *Geophysical Research Letters*, vol. 30, no. 11, 2003.

[31] R. Fovell and Y. Cao, "The santa ana winds of southern california: Winds, gusts, and the 2007 witch fire," *Wind and Structures An International Journal*, vol. 24, pp. 529–564, 06 2017. [Online]. Available: https://doi.org/10.12989/was.2017.24.6.529

[32] P. A. Sánchez-Pérez, M. Robles, and O. A. Jaramillo, "Real time markov chains: Wind states in anemometric data," *Journal of Renewable and Sustainable Energy*, vol. 8, no. 2, p. 023304, 3 2016.

[33] S. Carreón-Sierra, A. Salcido, T. Castro, and A.-T. Celada-Murillo, "Cluster analysis of the wind events and seasonal wind circulation patterns in the mexico city region," *Atmosphere*, vol. 6, no. 8, pp. 1006–1031, 2015.

[34] L. Dong, L. Wang, S. F. Khahro, S. Gao, and X. Liao, "Wind power day-ahead prediction with cluster analysis of nwp," *Renewable and Sustainable Energy Reviews*, vol. 60, pp. 1206–1212, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032116001362

[35] A. Trasviña, M. Ortiz-Figueroa, H. Herrera, M. A. Cosío, and E. González, "'Santa Ana winds' and upwelling filaments off northern Baja California," *Dynamics of Atmospheres and Oceans*, vol. 37, no. 2, pp. 113–129, 2003. [Online]. Available: http://dx.doi.org/10.1016/S0377-0265(03)00018-6

[36] J. O. I. Castro, A. A. Ramirez, M. Z. Machado, M. A. Vazquez, and N. L. Osuna, "Wind states and power curve modeling: A case study for la rumorosa i wind farm," *Wind and Structures*, vol. 39, no. 3, pp. 163–174, 2024.

[37] P. Kaufmann and C. D. Whiteman, "Cluster-analysis classification of wintertime wind patterns in the grand canyon region," *Journal of Applied Meteorology*, vol. 38, no. 8, pp. 1131–1147, 1999.

[38] S. Konishi and G. Kitagawa, *Information Criteria and Statistical Modeling*, 1st ed. Springer Publishing Company, Incorporated, 2007.

[39] G. Claeskens and N. L. Hjort, "The Bayesian information criterion," in *Model Selection and Model Averaging*. Cambridge: Cambridge University Press, 2001, pp. 70–98.

[40] S. Conil and A. Hall, "Local regimes of atmospheric variability: A case study of southern california," *Journal of Climate*, vol. 19, no. 17, pp. 4308–4325, 2006. [Online]. Available: 10.1175/JCLI3837.1

[41] R. Castro, A. Parés-Sierra, and S. G. Marinone, "Evolution and extension of the santa ana winds of february 2002 over the ocean, off california and the baja california peninsula," *Ciencias Marinas*, vol. 29, no. 3, pp. 275–281, 2003. [Online]. Available: http://www.cienciasmarinas.com.mx/index.php/cmarinas/article/view/158

[42] M. Arellano, C. Minutti, and A. Huerta, "WindC1IA," 08 2022. [Online]. Available: https://github.com/estudiovientos/saw/tree/Versi%C3%B3n-2

Magali Arellano-Vázquez holds a Ph.D. in Computer Science from the National Autonomous University of Mexico (UNAM). She is a researcher at INFOTEC, Center for Research and Innovation in Information and Communication Technologies, at the National Laboratory of Future Internet in Aguascalientes, Mexico. Her research interests include distributed computing, wind energy, complex systems, temporal networks, data analysis, and clustering algorithms.

Marlene Zamora-Machado holds a Ph.D. in Engineering from the Autonomous University of Baja California. She is a researcher at the Faculty of Engineering in the Renewable Energy Program in Mexicali. Her research focuses on the energy contribution of various meteorological phenomena to wind turbines. Her research interests include wind energy production, wind dynamics, and renewable energy sources.

Miguel Robles-Pérez holds a Ph.D. from the Autonomous University of the State of Morelos. He completed a two-year postdoctoral stay at Aalto University of Technology and Design in Helsinki, Finland. He is currently a researcher at the Institute of Renewable Energies (IER). He has published 32 articles in JCR-indexed journals. His research interests include statistical mechanics and materials science. His work in liquid theory has contributed to the study of the thermodynamics of simple liquids and their phase transitions. He is currently involved in multidisciplinary, multiscale modeling projects for materials used in energy conversion and storage devices. He also applies stochastic models and data science techniques to wind dynamics and complex systems related to renewable energy sources.

Carlos Minutti-Martinez holds a Ph.D. in Computer Science from the National Autonomous University of Mexico (UNAM). He completed a research stay at the University of Waterloo, Canada, under the Emerging Leaders in the Americas Program. His expertise in data science includes consulting and working as a research associate for the Consortium of Artificial Intelligence. He has completed postdoctoral fellowships at the National Polytechnic Institute and UNAM. He currently serves as a Research Scientist at INFOTEC in Mexico.

Oscar Alfredo Jaramillo holds a Ph.D. in Mechanical Engineering with *Magna Cum Laude* honors from the National Autonomous University of Mexico (UNAM). He is currently a Principal Researcher "B" at UNAM. Over the past five years, he has published 17 articles in JCR-indexed journals. His research interests include concentrated solar energy, wind energy systems for urban applications, entropy and energy analysis in renewable energy systems, and the integration of renewable energy sources.