symmetry

MDPI

Article

SICABI: Symmetry-Informed Stochastic Modeling via Dominant-Period Stationarity and Recursive Adaptive Parametric Density Estimation

Daniel Canton-Enriquez, Jorge-Luis Perez-Ramos, Selene Ramirez-Rosales, Luis-Antonio Diaz-Jimenez, Ana-Marcela Herrera-Navarro and Hugo Jimenez-Hernandez *

Facultad de Informática, Universidad Autónoma de Queretaro, Av. de las Ciencias S/N, Juriquilla, Santiago de Queretaro 76230, Mexico; daniel.canton@uaq.mx (D.C.-E.); jorge.luis.perez@uaq.edu.mx (J.-L.P.-R.); selene.ramirez@uaq.mx (S.R.-R.); luis.diaz@uaq.mx (L.-A.D.-J.); mherrera@uaq.mx (A.-M.H.-N.)
* Correspondence: hugo.jimenez@uaq.mx; Tel.: +52-442-473-2064

Abstract

Wind dynamics in urban environments exhibit non-stationarity and marked spatial variability, complicating stochastic modeling when a single global distribution is assumed. This article discusses the estimation of wind density under quasi-stationary regimes at the local level using SICABI, a two-phase framework: (i) Stationary Region Identification (ISR) estimates, through spectral power analysis, a specific dominant period for each location and validates the induced subsampling using the Augmented Dickey–Fuller (ADF) test, and (ii) RAPID adjusts an adaptive parametric density by recursively updating the mixture parameters and creating new components when a normalized membership distance exceeds a threshold. The analysis uses wind speed records collected from eight stations in the Metropolitan Area of Queretaro, Mexico, during the period from 1 January 2023 to 31 December 2023, aggregated at a 10 min resolution, from which $X_{\delta,s}$ is constructed for each site. RAPID is compared against Gaussian Kernel Density Estimation (KDE) with Silverman bandwidth and EM-fitted Gaussian mixtures with BIC-based selection ($K_{\max} = 12$). The resulting densities were compared with an empirical density estimated from a histogram over a fixed grid ($m = 50$) using the MISE and RMSE metrics. The results reveal marked site-dependent differences in dominant periodicity and residual behavior, including asymmetry and heavy tails. ISR identified dominant periods ranging from 37 to 166 days, and RAPID adapted its complexity with $K_s \in [5, 10]$ without fixing the number of mixture components in advance. Quantitatively, RAPID achieved the lowest RMSE at 6/8 sites and the lowest MISE at 5/8 sites, while also exhibiting shorter execution times than KDE and MoG under the same input $X_{\delta,s}$. The results support RAPID as a competitive adaptive method for site-specific density estimation in non-stationary urban climate signals. In this context, local regimes can be viewed as approximate invariants under time translation in the weak stochastic sense, while deviations from this assumption are reflected in increased distributional complexity across sites.

Keywords: local stationarity; spectral analysis; adaptive mixture models; wind speed; pattern recognition; symmetry-aware algorithms

Check for updates

Academic Editors: Zeev Volkovich, Renata Avros, Dvora Toledano-Kitai and Jie Yang

Received: 6 March 2026
Revised: 8 April 2026
Accepted: 17 April 2026
Published: 20 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Atmospheric models use historical data to characterize climate dynamics and to understand the interactions among meteorological variables. Among these variables,

wind speed is especially relevant because its variations affect others such as temperature, atmospheric pressure, and humidity [1]. These variables are important for public policy and decision-making, particularly in the health sector, where significant links between extreme weather conditions and increased mortality and morbidity have been documented [2]. Likewise, climate behavior affects key economic sectors such as agriculture [3], water management [4], tourism [5], and energy production [6,7], all of which depend strongly on environmental conditions for planning and operation.

Climate change is a reality that makes it difficult to estimate climate variables, such as the intensity of rainfall and drought periods, as well as the occurrence of extreme events, such as hurricanes, heat waves, and forest fires, which demand automated decision-support tools [8,9]. One of the main challenges in climate prediction is modelling the stochastic nature of climate variables, which affects model accuracy and, consequently, decision-making. Likewise, deterministic models struggle to represent the dynamics and temporal variability of wind speed [10]. Moreover, local topographic conditions introduce additional variations, further complicating modeling.

The temporal variation in climate variables leads to non-stationary, quasi-periodic behavior over extended time intervals [11]. Nonetheless, these quasi-periodic patterns permit local approximations by assuming stationarity over shorter time intervals [7,12,13]. Local stationarity enables the use of stationary statistical methods to obtain reasonable predictions and detect useful trends [14]. However, treating wind speed as a complex environmental variable introduces an additional challenge: efficiently estimating its PDF [15]. In this context, probabilistic models provide a natural alternative because they allow the data to be represented through parametric PDFs and facilitate the identification of stochastic structures in the observations.

The methodological objective of this study is not stochastic modeling in general, but rather the estimation of PDFs at the local scale for each site, accounting for non-stationary, spatially heterogeneous wind conditions. In urban areas, measurements of climatic variables such as wind are influenced by local exposure, building topography, and terrain. This can cause their statistical behavior to vary considerably from one site to another and over time. Under these conditions, a single global density or a fixed mixture of components may not adequately reflect the following challenges: local tails, secondary modes, or changes in the uncertainty structure. This motivates a framework that first identifies a locally quasi-stationary regime and then estimates the corresponding density through an adaptive parametric structure.

The current approaches for estimating density in environmental series can be broadly grouped into three categories. First of all, global parameter models adopt a predetermined functional shape and, frequently, a fixed set of components, which may be unduly restrictive in heterogeneous local regulations. Second, non-parametric estimators such as KDE offer flexibility, but their performance depends heavily on bandwidth selection and may oversmooth multimodal or heavy-tailed distributions. Third, methods based on local stationarity provide a principled way to isolate intervals with approximately stable statistics, but, by themselves, they do not define how the resulting densities should be estimated adaptively. The gap addressed in this work lies in combining local regime identification with an adaptive parametric density estimator that does not require the number of components to be fixed in advance.

The automatic construction of stochastic models involves designing probabilistic models whose structure and parameters are inferred from data. This approach minimizes human intervention and avoids imposing rigid global assumptions. Whenever new observations are added, the model is updated iteratively, allowing it to adapt to the local patterns, temporal variability, and changes in the dynamics of the underlying process [16,17]. In

urban climate applications, local wind behavior may vary substantially from one site to another due to topography, land cover, and exposure conditions, which justifies the use of site-specific stochastic modeling strategies [18,19]. In this context, the automatic construction of a stochastic model from historical data should consider three stages: (a) identifying local stationarity; (b) estimating the parameters of the PDF through parametric structures; and (c) validating the quality of the resulting fit.

First, (a) identifying local stationarity aims to find intervals where the statistical properties remain roughly constant. This enables the use of segment-wise stationary models [12,14]. To estimate a dominant periodicity, spectral analysis is employed, and the induced subsampling is then statistically validated using the Augmented Dickey–Fuller (ADF) test [20,21]. The corresponding fundamentals and formal definitions are presented in Section 2. Subsequently, (b) the parameters of the PDF are estimated using a parametric mixture model. This approach prioritizes recursive updates, enabling incremental adjustments based on identified segments [16,17]. Finally, (c) the fit is validated through discrepancy criteria and residual analysis, and is further supported by comparisons with established reference methods.

The Metropolitan Area of Queretaro (MAQ), Mexico (see Figure 1), is used as a case study for two main reasons: (i) the need to characterize the local variability of meteorological variables in urban environments for improved monitoring and decision-making, and (ii) the availability of high-frequency and multi-site measurements from a local meteorological network. This motivation is consistent with international guidelines for integrated urban services and urban climate monitoring, which emphasize the value of high-resolution local observations in cities [22,23]. In particular, this study uses the RedCIAQ-UAQ network, which provides minute-by-minute climatological data and constitutes a suitable source for evaluating automated methods based on temporal segmentation and density estimation [24].

This article presents SICABI (in the Zapotec language of the Isthmus of Tehuantepec, “sicabi” means “like the wind,” symbolizing the dynamic and adaptive nature of the proposed framework) (Stochastic Inference of Computationally Adaptive Behavioral Structures), a framework for the automatic construction of stochastic models from raw data acquired at MAQ weather stations. The dataset includes geographic information for each site, together with minute-by-minute measurements collected at eight stations from 1 January 2023 to 31 December 2023. The climatic records were aggregated every 10 min, and the model identifies local stationarity periods at each geographic location. The framework then estimates the corresponding density using a combination of parametric functions via a recursive version of the Expectation-Maximization (EM) algorithm. In this way, the resulting model changes according to the prevailing local regime at each site, which is conditioned by topography and wind dynamics.

In light of the above, the main contributions of this work are as follows:

- SICABI, a two-stage framework for stochastic density estimation under locally quasi-stationary regimes in environmental time series.
- ISR (Identification of Stationary Regions), a procedure that identifies a dominant local period through spectral analysis and validates the induced subsampling using the ADF test.
- RAPID (Recursive Adaptive Parametric Density Estimation), an adaptive density estimator that recursively updates mixture parameters and creates new components when the local membership criterion is not satisfied.
- An empirical evaluation against KDE and MoG under a common protocol, illustrating the trade-off between adaptive complexity, fitting accuracy, and computational cost in an urban climate case study.

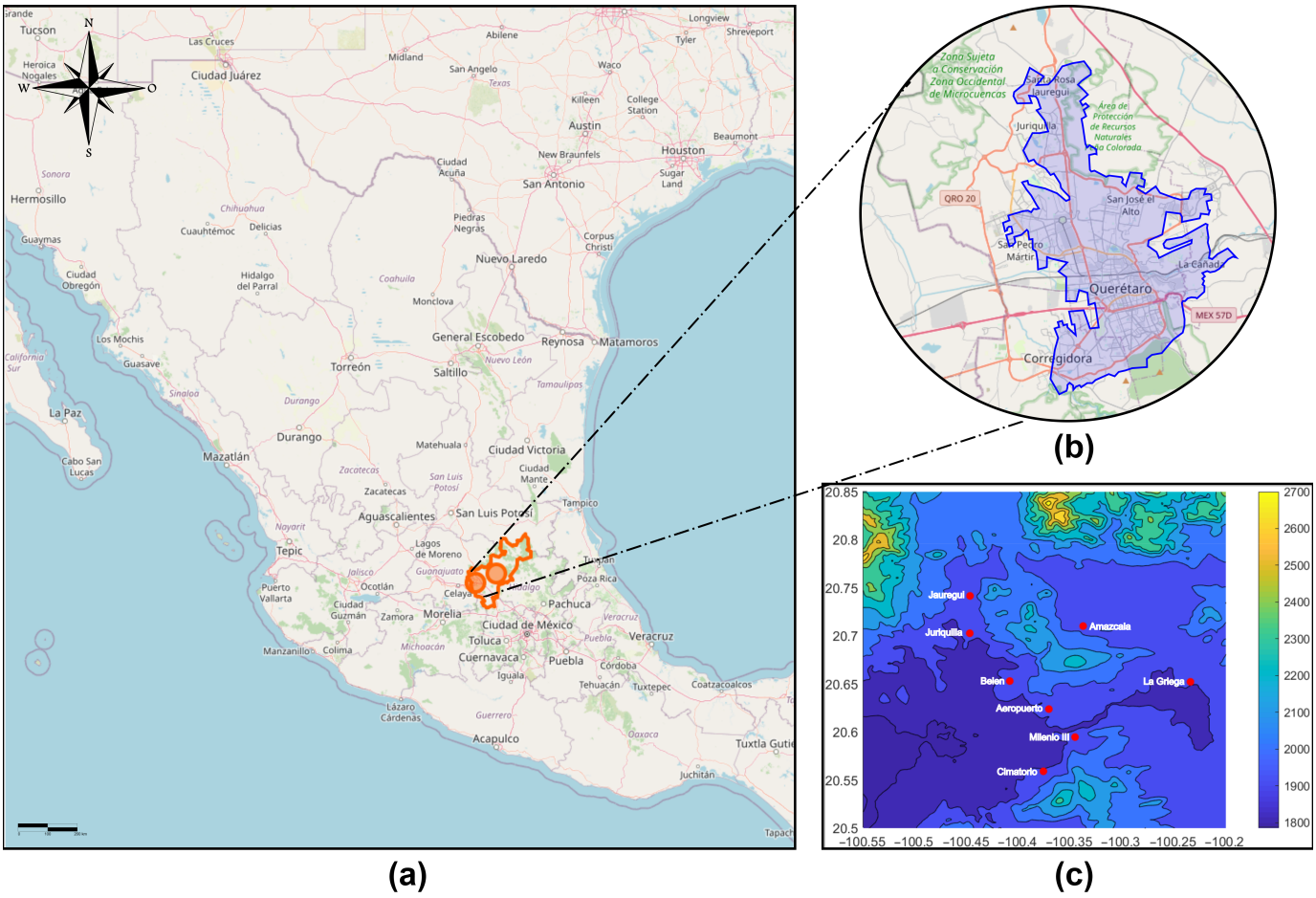


Figure 1. Geographical location of the case study and distribution of stations. (**a**) Location of the state of Queretaro (orange outline) within Mexico. (**b**) Position of the Metropolitan Area of Queretaro within the state (blue outline). (**c**) Representation of the MAQ in relief, with a color palette (elevation) and the spatial distribution of the eight weather stations considered (red circles).

2. Theoretical Foundations

The fundamentals of the framework are described in this section. These are divided into three stages: (i) time segmentation and local stationarity, (ii) density estimation using parametric mixtures, and (iii) an adaptive density model based on data flow.

2.1. Identification of Stationary Regions

This section introduces the ISR (Identification of Stationary Regions), as climate series often exhibit non-stationary behavior over long horizons; however, they can be approximated as locally stationary over shorter intervals [12,14]. Under this hypothesis, identifying stationary regions involves finding segments where statistical properties remain relatively constant.

Spectral analysis helps characterize quasiperiodic patterns and regime changes. Let $x[n]$ be a discrete series of length N. The Discrete Fourier Transform (DFT) [25] is defined as

$$X[k] = \sum_{n=0}^{N-1} x[n]\, e^{\frac{-2\pi i k n}{N}}, \quad k = 0, 1, \ldots, N-1 \tag{1}$$

where it have a computational complexity of $\mathcal{O}(N^2)$. To address this limitation for long time series, the Fast Fourier Transform (FFT) [26] enables efficient computation of the DFT, reducing the complexity to $\mathcal{O}(N \log N)$ using the Cooley-Tukey divide-and-conquer scheme [21].

In the frequency domain, the spectral power of a signal quantifies the energy contribution of each frequency component, which becomes defined as

$$P[k] = \frac{|X[k]|^2}{N}, \quad k = 0, 1, \ldots, N-1, \tag{2}$$

where the expression allows identification of dominant frequencies that represent periodic structures. Consequently, a time region is considered a candidate for stationarity if its spectral signature remains stable over time (e.g., under a sliding window approach).

Once the dominant frequencies become identified, a validation test is usually supported by unit root tests such as the ADF [20] (and, complementarily, the KPSS [27]), to confirm that the identified segments meet stationarity criteria from an inferential perspective [28].

2.2. Mixture of Parametric Functions

Density estimation via a mixture of parametric functions yields a model based on a complex distribution, typically a convex combination of components from a known family of distributions. Without loss of generality, the Mixture of Gaussians (MoG) approach has been used in several contexts. Consider a univariate continuous random variable x, a MoG with K components can be expressed as follows as PDF approximation:

$$f(x) = \sum_{i=1}^{K} \phi_i \mathcal{N}\left(x \mid \mu_i, \sigma_i^2\right) \tag{3}$$

where $\phi_i \in [0, 1]$ denotes the probability of observation.

The enumeration of K elements expressed each element i by an individual tuple $\left\{\phi_i, \mu_i, \sigma_i^2\right\}_{i=1}$, the parameters of each parametric function that might be computed iteratively with Expectation-Maximization (EM), which aims to maximize the model's log-likelihood [16,17,29]. In sequential (data-flow) scenarios, online variants of EM exist that update parameters as new observations arrive; these are typically used in classic adaptive modeling applications [30].

A structural limitation of (3) is that it fixes K a priori. Information criteria such as AIC/BIC are used to select K; however, in contexts with non-stationarity or changing multimodality, a scheme with dynamic component growth may be preferable.

2.3. RAPID: Recursive Adaptive Parametric Density Estimation

This work introduces an adaptive density model that builds on the concept of parametric mixtures. This model offers two key benefits: (i) it facilitates the online updating of important parameters, and (ii) it allows for an increase in the number of components in response to changes in the observed structures.

Let $X = \{x_1, x_2, \ldots, x_t\}$ be an observed sequence of a climatic variable at any geographic location, where the density is represented as the following mixture of Gaussians (MoG):

$$f(x) = \sum_{i=1}^{K} \phi_i \mathcal{N}\left(x \mid \mu_i, \sigma_i^2\right) \tag{4}$$

where the parameter K can change over time. To assign a new observation x to an existing component, a normalized distance-based approach is utilized, as shown below:

$$d_i(x) = \frac{(x - \mu_i)^2}{\sigma_i^2} \tag{5}$$

When $\min_i d_i(x)$ exceeds a defined threshold, denoted as λ, a new component is created; otherwise, the parameters of the selected component are updated recursively. A common update method compatible with recursive approaches, such as those used in EM techniques, is outlined in [31]:

$$\mu_{k,t+1} = \rho\, \mu_{k,t} + (1-\rho)\, x_t, \tag{6}$$

$$\sigma^2_{k,t+1} = \rho\, \sigma^2_{k,t} + (1-\rho)\, (x_t - \mu_{k,t+1})^2. \tag{7}$$

where $\rho \in [0,1]$ is a parameter that allows control of the permissive adaptability rate of the model.

3. Materials and Methods

This section describes the operational methodology of the SICABI framework and its computational implementation. The identification of local stationarity is based on Section 2.1. In contrast, the formulation of the parametric mixture and the adaptive scheme are based on Sections 2.2 and 2.3. The data source, the general processing flow, the implemented algorithms, and their computational complexity are detailed below.

To facilitate the reading of the methodological sequence, Figure 2 summarizes the operational workflow of the proposed framework. The process begins with the acquisition and regularization of meteorological records, continues with the identification of locally quasi-stationary regimes through ISR, and concludes with adaptive density estimation through RAPID and its comparison against reference methods.

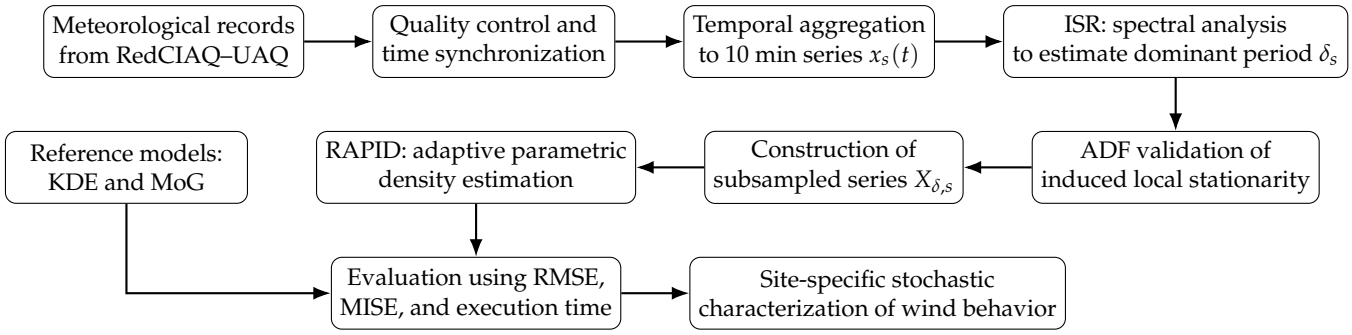


Figure 2. General workflow of SICABI, from meteorological data preprocessing and local stationarity identification to adaptive density estimation and performance evaluation.

3.1. Study Area and Data Description

The Metropolitan Area of Queretaro (MAQ), Mexico, was selected as the study area due to its urban and peri-urban heterogeneity, which affects local atmospheric behavior. The MAQ is located in the Queretaro Valley and its geographical transition zones, including the Buenavista Valley to the north and the Amazcala Valley to the east. These differences in topography, exposure, and local relief contribute to spatial variability in wind dynamics, even among relatively nearby sites. Such heterogeneity is especially relevant in urban environmental monitoring, where wind influences ventilation, gas transport, and pollutant dispersion [32]. Under these conditions, location-specific probabilistic schemes are more appropriate than uniform global assumptions [22,23].

Historical records from automatic weather stations in the RedCIAQ–UAQ network were used to construct the stochastic model. The network provides real-time meteorological observations at minute-level resolution [24,33]. In this study, wind speed was selected as the target variable. For spatial contextualization, two complementary sources of information were considered: (i) wind-speed time series by site, and (ii) static geographic metadata for each station, including latitude, longitude, and elevation above mean sea level.

Figure 1 presents the geographical location of the study area and the spatial distribution of the eight monitored stations. Table 1 summarizes the geographic attributes of each site together with descriptive statistics of wind speed. This organization allows the spatial context, the monitored locations, and the observed variability to be interpreted jointly.

Table 1. Geographic information of the monitored sites and descriptive statistics of wind speed. Note: Meteorological observations were obtained from RedCIAQ–UAQ. Latitude, longitude, and elevation are static geographic descriptors used for spatial contextualization, whereas the wind-speed statistics were computed from the observation period analyzed in this study.

Site	Latitude (North)	Longitude (West)	Elevation (m.a.s.l.)	Wind Speed (m/s)		
				Maximum	Average	SD
Belen (BN)	20°39′31″	100°24′31″	1964	11.1	0.857	1.307
Amazcala (AZ)	20°42′37″	100°20′16″	1970	16.1	0.438	1.318
Aeropuerto (AP)	20°37′26″	100°22′15″	1973	20.0	4.323	2.388
La Griega (LG)	20°33′32″	100°22′33″	1909	22.0	1.061	1.697
Cimatario (CT)	20°33′33″	100°22′34″	1924	14.7	1.366	1.562
Jauregui (JR)	20°44′30″	100°26′49″	1968	13.1	1.308	1.453
Milenio III (M3)	20°35′41″	100°20′44″	1966	12.5	1.407	1.393
Juriquilla (JQ)	20°42′10″	100°26′50″	1941	13.3	2.853	1.572

Table 1 reveals marked spatial heterogeneity in wind behavior across the monitored sites. Aeropuerto exhibits the highest average wind speed (4.323 m/s), suggesting a more persistent flow regime, whereas Amazcala shows the lowest average value (0.438 m/s), consistent with weaker local circulation. La Griega records the highest maximum wind speed (22.0 m/s), indicating the occurrence of stronger extreme events, while variability also differs substantially among sites, with standard deviations ranging from 1.307 to 2.388 m/s. These differences support the use of site-specific stochastic modeling strategies instead of a single homogeneous probabilistic representation for the entire MAQ.

The apparent temporal distinction between the geographic descriptors and the meteorological statistics is due to their different roles in the analysis. Latitude, longitude, and elevation are fixed attributes of each station, whereas the wind-speed statistics correspond to the observation period considered in this study, namely from 1 January 2023 to 31 December 2023. This distinction was made explicit to avoid confusion between static spatial metadata and time-dependent meteorological measurements.

Operationally, the records were organized into standardized time series by site, denoted by $x_s(t)$. Before modeling, basic quality-control procedures were applied, including timestamp synchronization, duplicate removal, and treatment of missing values by linear interpolation. To homogenize the analysis and reduce high-frequency variability, the minute-level observations were aggregated to a 10 min resolution using the operator $\mathcal{A}(\cdot)$, defined as the mean over non-overlapping windows. This produced a regular sequence $x_s(t)$ for each site s, which was used as input to the ISR stage shown in Figure 3.

3.2. General Scheme of the SICABI Framework

Figure 3 summarizes the operational flow of the SICABI framework. Conceptually, the assessment of local stationarity is based on Section 2.1. On the other hand, the formulation of the parametric mixture and the adaptive scheme is based on Sections 2.2 and 2.3, respectively. This framework summarizes the operational process that helps discover, in non-stationary scenarios, local stationarity, enabling the creation of a stochastic model locally explained by parametric approaches.

This chapter describes the specific parameterization and computational implementation that materialize these foundations.

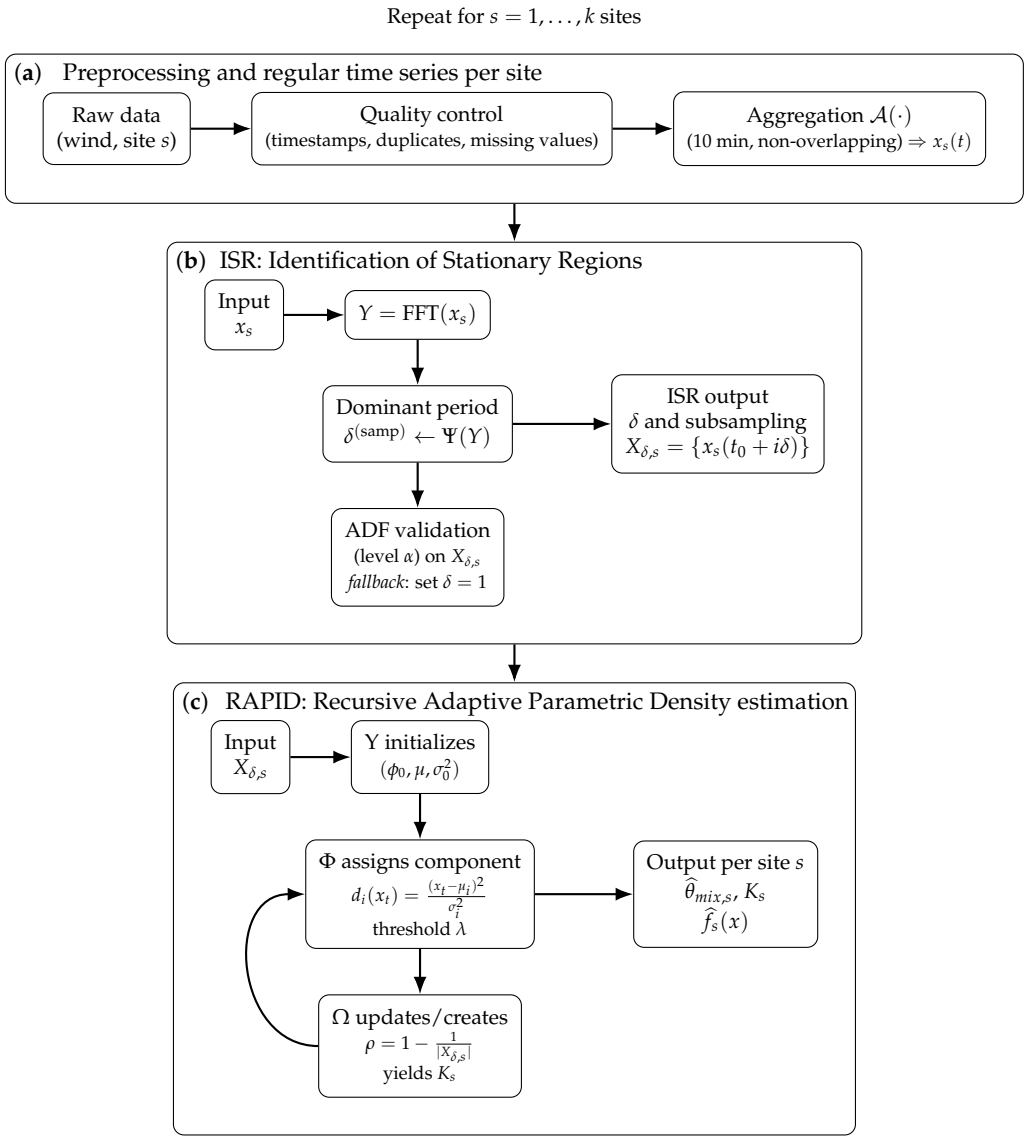


Figure 3. SICABI block diagram in three stages: (**a**) preprocessing and construction of the regular series $x_s(t)$ per site; (**b**) ISR estimates the dominant regime δ via FFT, spectral power (excluding DC), and ADF validation, defining the subsampling $X_{\delta,s}$; (**c**) RAPID fits an adaptive parametric mixture from $X_{\delta,s}$ using operators Υ, Φ, Ω (controlled by $\lambda, \rho, \sigma_0^2, \phi_0$), producing $\widehat{f}_s(x)$, $\widehat{\theta}_{mix,s}$, and K_s.

3.3. Algorithm Implementation

The SICABI framework implementation consists of two main steps: ISR and RAPID (Figure 3). ISR performs local stationarity identification using spectral analysis and inferential validation (see Section 2.1). Subsequently, RAPID constructs a density estimate based on a mixture of parametric functions with recursive updating and dynamic growth of the number of components (see Sections 2.2 and 2.3). This section details the parameterization decisions, the algorithmic structure, and the computational implementation aspects. The implemented versions ISR, and RAPID are presented below (see Algorithms 1 and 2, respectively).

Algorithm 1 Identification of Stationary Regions (ISR)

Require: x ▷ Historical data of length N.
Ensure: δ ▷ Estimated dominant stationarity period (or -1).
procedure ISR(x)
  $Y \leftarrow \mathrm{FFT}(x)$ ▷ Fast Fourier Transform
  $\delta \leftarrow \Psi(Y)$ ▷ Dominant period from spectral power
  $z \leftarrow \textsc{Subsample}(x, \delta)$ ▷ $z = \{x_1, x_{1+\delta}, x_{1+2\delta}, \dots\}$
  $p \leftarrow \mathrm{ADFTEST}(z)$
  if $p < \alpha$ **then**
    return δ
  else
    return -1 ▷ Fallback policy described in text (Section 3.4/below)
  end if
end procedure

Algorithm 2 Recursive Adaptive Parametric Density estimation (RAPID)

Require: $X_{\delta,s}$ ▷ Subseries sampled every δ units.
Require: λ ▷ Belonging threshold.
Ensure: $\widehat{\theta}_{mix}$ ▷ Mixture parameters $(\phi_i, \mu_i, \sigma_i^2)$.
procedure RAPID($X_{\delta,s}, \lambda$)
  Initialize $\widehat{\theta} \leftarrow \Upsilon(\phi_0, x_1, \sigma_0^2)$ ▷ First component
  $\rho \leftarrow 1 - \frac{1}{|X_{\delta,s}|}$ ▷ Learning factor
  for all $x_t \in X_{\delta,s}$ **do**
    $(i^*, d_{\min}) \leftarrow \Phi(\widehat{\theta}, x_t)$ ▷ Nearest component and distance
    if $d_{\min} \leq \lambda$ **then**
      $\widehat{\theta} \leftarrow \Omega(\widehat{\theta}, i^*, x_t, \rho)$ ▷ Update (ϕ, μ, σ^2)
    else
      $\widehat{\theta} \leftarrow \widehat{\theta} \cup \Upsilon(\phi_0, x_t, \sigma_0^2)$ ▷ New component
    end if
  end for
  return $\widehat{\theta}_{mix}$
end procedure

Once the dominant periodicity is identified and the subsampling's stationarity is validated using ISR, RAPID fits an adaptive parametric mixture. The implementation employs three operators: Υ (initialization), Φ (assignment/membership criterion), and Ω (recursive parameter update), in accordance with the theoretical formulation of Section 2.3.

Algorithm 1 assumes that the data corresponds to complete observations, considering the level of noise inherent in the acquisition and/or transmission process. For the estimation of stationarity, the operator Ψ calculates the spectral power over the positive frequencies and determines the dominant period δ according to Equation (8):

$$\delta \leftarrow \mathrm{period}[\arg\max(P)] \tag{8}$$

The spectral power $Y = \mathrm{FFT}(x)$ and $P[k]$ defined in Equation (2). The operator $\Psi(\cdot)$ calculates the dominant period excluding the DC component ($k = 0$) and limiting it to positive frequencies:

$$k^* \leftarrow \underbrace{\arg\max(P[k])}_{k \in \{1, \dots, \lfloor N/2 \rfloor\}} \qquad \delta \leftarrow \left\lceil \frac{N}{k^*} \right\rceil \tag{9}$$

Equation (9), k^*, corresponds to the index of the frequency with the highest energy (other than DC), and δ represents the dominant periodicity in the number of samples. This

estimate is treated as a sampling candidate and subsequently validated inferentially using ADF in ISR (Section 2.1).

Regarding RAPID, the Y operator initiates a component with a mean of $\mu = x_1$ and an initial variance of σ_0^2. The Φ operator determines whether x_t belongs to the set using the threshold λ, and returns the nearest component i^* along with the minimum distance $d_{\min}$. Finally, Ω recursively updates the parameters of the selected component, according to the learning rate ρ (Section 2.3).

3.4. Configuration and Hyperparameters

Table 2 summarizes the hyperparameters set for ISR and RAPID. The same applies to the comparison protocol (KDE/MoG). To ensure reproducibility and comparability, these values were kept unchanged at all sites; in contrast, site-dependent outputs (e.g., δ and K_s) are reported in Section 4.3.

Table 2. Configuration and hyperparameters used in ISR, RAPID, and the comparison protocol.

Parameter	Description	Value	Use
Δt	Resolution after aggregation ($\mathcal{A}(\cdot)$)	10 min	Series $x_s(t)$
α	Significance level (ADF)	0.05	ISR validation
δ	Dominant subsampling step	ISR	Construction $X_{\delta,s}$
λ	Membership threshold (normalized distance)	3	Creation of RAPID components
ρ	Learning rate (recursive update)	$\frac{\lvert X_{\delta,s}\rvert - 1}{\lvert X_{\delta,s}\rvert}$	Update Ω
σ_0^2	Initial variance of new component	1	Initialization Y
ϕ_0	Initial weight of new component	1	Initialization Y
m	Mesh size G for evaluating densities	50	RMSE/MISE
$K_{\max}$	Max. components evaluated in MoG (BIC)	12	MoG comparison
R	Repetitions for execution times	30	Time table

If the ADF test applied to $X_{\delta,s}$ does not reject the null hypothesis of unit root (i.e., $p \geq \alpha$), alternative spectral peaks are not explored. Instead, $\delta = 1$ (no subsampling) is set, and RAPID is continued, recording this case for analysis in the results section.

3.5. Implementation Issues

This subsection discusses the computational complexity of the three implemented algorithms. The complexity is addressed in three main stages:

1. Spectral segmentation using FFT;
2. Statistical validation of stationarity and periodicity selection in ISR;
3. Density estimation with adaptive growth in RAPID.

Hereafter, N denotes the length of the aggregated series $x_s(t)$ per site. Let $T(N)$ be the time required to compute the FFT of a series of size N. The recurrence of the Cooley-Tukey scheme is $T(N) = 2T(N/2) + \mathcal{O}(N)$, which is solved as $T(N) = \mathcal{O}(N \log N)$. This complexity improves upon direct DFT calculations, which have a cost of $\mathcal{O}(N^2)$.

ISR applies FFT with a complexity of $\mathcal{O}(N \log N)$. Estimating the dominant period $\Psi(Y)$, which inspects spectral magnitudes, requires $\mathcal{O}(N)$. Subsampling every δ units consumes $\mathcal{O}(N/\delta)$, and in the worst case (when $\delta = 1$) it is $\mathcal{O}(N)$. The ADF test, for a fixed number of delays, is typically considered to be $\mathcal{O}(N)$. Consequently, ISR is FFT-dominated:

$$\mathcal{O}(N \log N) + \mathcal{O}(N) = \mathcal{O}(N \log N) \tag{10}$$

Let $X_{\delta,s}$ be the subseries (per site s) obtained by the subsampling induced by δ, and let

$$M \leftarrow |X_{\delta,s}| \tag{11}$$

its length. Under regular subsampling conditions, $M \approx N/\delta$, where N denotes the length of the aggregated series $x_s(t)$.

The initialization of ρ has a constant complexity. The main cycle records M observations. In these cases, if the number of components increases linearly with t, the evaluation continues with the loss criterion at $\mathcal{O}(M^2)$. In typical situations where the number of components is stabilized in $K \ll M$, the estimate of the computer cost is calculated in $\mathcal{O}(KM)$.

Therefore, the worst-case computational complexity of RAPID is defined as

$$\mathcal{O}(M^2) = \mathcal{O}\left(|X_{\delta,s}|^2\right) \approx \mathcal{O}\left(\frac{N^2}{\delta^2}\right) \tag{12}$$

In the usual form where K stabilizes and δ is bounded, the cost $\mathcal{O}(KM) = \mathcal{O}(K|X_{\delta,s}|)$ is nearly linear in N (provided that K is bounded).

ISR has dominant complexity $\mathcal{O}(N \log N)$. Let $M = |X_{\delta,s}|$ be the length of the subseries, with $M \approx N/\delta$ under regular subsampling. RAPID has a worst-case cost of $\mathcal{O}(M^2)$ and typically $\mathcal{O}(KM)$ when the number of components stabilizes at $K \ll M$.

The global cost is expressed as

$$\mathcal{O}\left(k\left(N \log N + M^2\right)\right) \quad \text{(worst case)}, \tag{13}$$

and, in the typical regime with a bounded value of K, a final complexity is obtained, described as

$$\mathcal{O}(k(N \log N + KM)) \tag{14}$$

Although the worst theoretical case for RAPID is $\mathcal{O}(M^2)$ when the number of components grows without bound, in the configuration used ($\lambda = 3$), the number of components stabilizes at a small value $K_s \ll M$ (see Table 3). In this typical regime, the cost of the density estimation phase in RAPID is approximated by $\mathcal{O}(K_s M)$, and when considering K_s bounded (e.g., $K_s \in [5, 10]$ in our experiments), the complexity is effectively linear in the input size:

$$\mathcal{O}(K_s M) \approx \mathcal{O}(M), \text{ with } M = |X_{\delta,s}|. \tag{15}$$

This point is relevant for comparison, since ISR is applied as common preprocessing to construct $X_{\delta,s}$, and both KDE and MoG are fitted on the same subseries (Section 4.2). Consequently, the cost disparities observed in the experiments primarily reflect the complexity of the density phase of each estimator. In this phase, RAPID operates incrementally with an effective linear cost in M when K_s is restricted.

Table 3. Final number of components K_s generated by RAPID per site.

Sites	K_s
Belén	6
Amazcala	9
Aeropuerto	10
La Griega	8
Cimatario	5
Jáuregui	7
Milenio III	6
Juriquilla	10

4. Experimental Analysis and Results

This section details the experimental design used to analyze the performance of the SICABI framework and its ability to recognize stochastic structures in real meteorological data. First, (i) the predominant periodicity linked to local quasi-stationarity is examined using the ISR method. Then, (ii) the adaptive density estimate is obtained using the RAPID algorithm. Finally, (iii) a direct comparison is made with representative current-state techniques (KDE and MoG), following a uniform comparison protocol.

4.1. Experimental Process

The experimental process consists of the following stages:

1. Dominant Periodicity Estimation and Quasi-Stationality Validation. For each geographic site s, ISR is applied to the aggregate series $x_s(t)$ to estimate a dominant period and inferentially validate the quasi-stationarity of the associated subsampling.
2. Density Estimation with RAPID. From subsampling $X_{\delta,s}$ (defined by the dominant period), an adaptive parametric mixture is fitted using RAPID, evaluating its stability, efficiency, and fit to the empirical density.
3. Comparison with Alternative Methods. RAPID is compared with KDE and MoG using the same input set $X_{\delta,s}$, with quantitative metrics and graphical evidence. Considerations of computational complexity are also discussed.

4.2. Protocol and Metrics for Comparisons

The ISR algorithm initially estimates a dominant periodicity in the number of samples in the aggregated series, denoted by $\delta^{(\text{samp})} \in \mathbb{N}$, which is used as a subsampling step to construct

$$X_{\delta,s} = \{x_s(t_0), x_s(t_0 + \delta^{(\text{samp})}), x_s(t_0 + 2\delta^{(\text{samp})}), \dots\} \tag{16}$$

Since sampling occurs every $\Delta t = 10$ min, the period is reported in days for climatological interpretation:

$$\delta^{(\text{day})} = \delta^{(\text{samp})} \cdot \frac{10}{1440} = \frac{\delta^{(\text{samp})}}{144} \tag{17}$$

In this section, the values reported in tables correspond to $\delta^{(\text{day})}$.

To measure the discrepancy between the estimated density and the empirical reference, all functions were evaluated on a regular grid $G = \{g_j\}_{j=1}^{m}$ defined over the interval $[\min(X_{\delta,s}), \max(X_{\delta,s})]$, where g_j denotes the j-th grid point and m is the total number of grid points used in the discrete approximation. In this study, $m = 50$. A normalized histogram constructed from $X_{\delta,s}$ and evaluated on G was used to approximate the empirical density f_{emp}. The symbol $f(\cdot)$ denotes a modeled density estimated by RAPID, KDE, or MoG, whereas $f_{\text{emp}}(\cdot)$ denotes the empirical density. The quantity Δg represents the uniform spacing between adjacent grid points. Under this discrete representation, RMSE quantifies the average pointwise deviation over the grid, whereas MISE approximates the integrated quadratic error along the support. KDE and MoG were fitted on the same set $X_{\delta,s}$ used by RAPID at each site, so that the observed differences reflect the estimation mechanism rather than differences in preprocessing or data selection. In KDE, a Gaussian kernel was used, and the bandwidth was selected using Silverman's rule. In MoG, the number of components was selected using BIC, and the fit was performed using classical EM. To determine this, $K \in \{1, \dots, K_{\max}\}$ was evaluated, where $K_{\max} = 12$. The quantitative metrics reported are as follows:

$$\text{RMSE}(f, f_{\text{emp}}) = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(f(g_j) - f_{\text{emp}}(g_j)\right)^2} \tag{18}$$

$$\begin{aligned} \mathrm{MISE}(f, f_{\mathrm{emp}}) &= \sum_{j=1}^{m} \left(f(g_j) - f_{\mathrm{emp}}(g_j)\right)^2 \Delta g \\ \Delta g &= \frac{g_m - g_1}{m-1} \end{aligned} \tag{19}$$

4.3. Results

4.3.1. Estimation of Dominant Periodicity with ISR

In the first stage, ISR was used to examine the annual time series of each geographic MAQ site. The smoothed signal was spectrally estimated to reduce high-frequency noise (δ). The $X_{\delta,s}$ subsampling was validated using the ADF test.

Figure 4 illustrates the procedure for the Airport site. The top portion shows the original series and its smoothed version, while the bottom section displays the power spectrum. The dominant period is determined from the maximum spectral power (excluding DC), interpreted as the highest-energy cyclic component associated with the prevailing wind dynamics at the site.

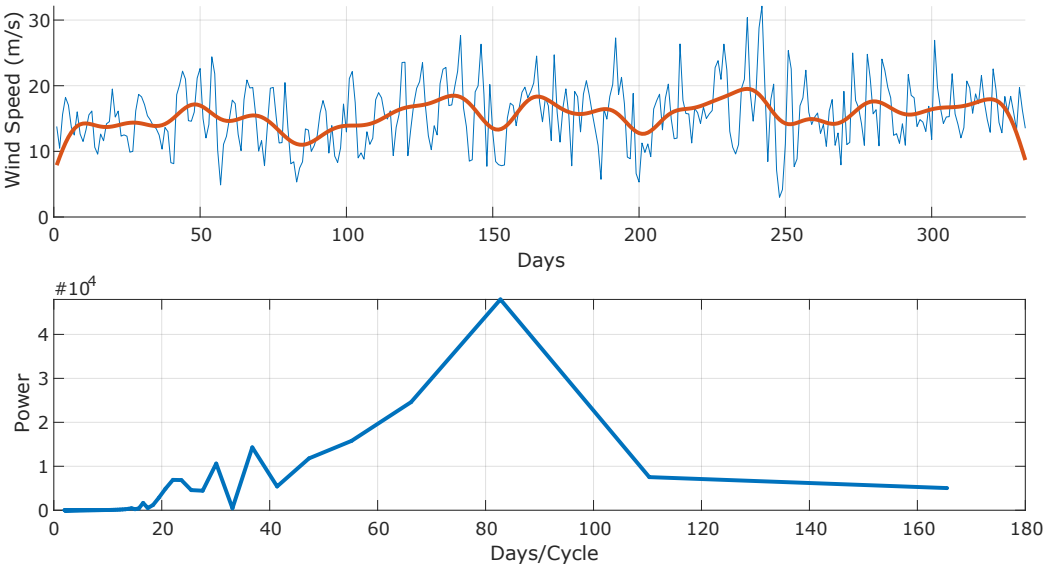


Figure 4. Example of spectral analysis to estimate the dominant wind periodicity (airport site). The upper figure shows the original data (blue line) and smoothed data (red line).

Table 4 shows the results for the dominant period $\delta^{(\mathrm{day})}$ for each site. Notable differences are observed across stations, suggesting that the wind's temporal structure is affected by local dynamics.

Table 4. ISR identifies the dominant period $\delta^{(\mathrm{day})}$ (rounded to the nearest integer).

Sites	$\delta^{(\mathbf{Day})}$
Belén	66
Amazcala	110
Aeropuerto	83
La Griega	110
Cimatario	166
Jáuregui	37
Milenio III	83
Juriquilla	83

As shown in Figure 1, the spatial distribution of stations located outside the valley (e.g., Amazcala and La Griega) and the Cimatario site exhibit longer dominant periods, while sites such as Jáuregui and Belén show shorter cycles. This variability is consistent with local conditions (microclimate, exposure, and topography) and supports the use of an adaptive estimator, such as RAPID, to capture site-dependent probabilistic structures.

4.3.2. Density Estimation with RAPID

Once the dominant periodicity per site was estimated, $X_{\delta,s}$ was constructed, and the PDF was adjusted using RAPID. This approach allows a mixture of parametric functions to be adjusted adaptively on a data set associated with a dominant temporal regime, avoiding biases characteristic of a global adjustment when temporal heterogeneity is present.

Figure 5 shows a representative case of the adjustment obtained. It can be seen that RAPID reproduces the general shape of the density and local modes without requiring a predefined number of components.

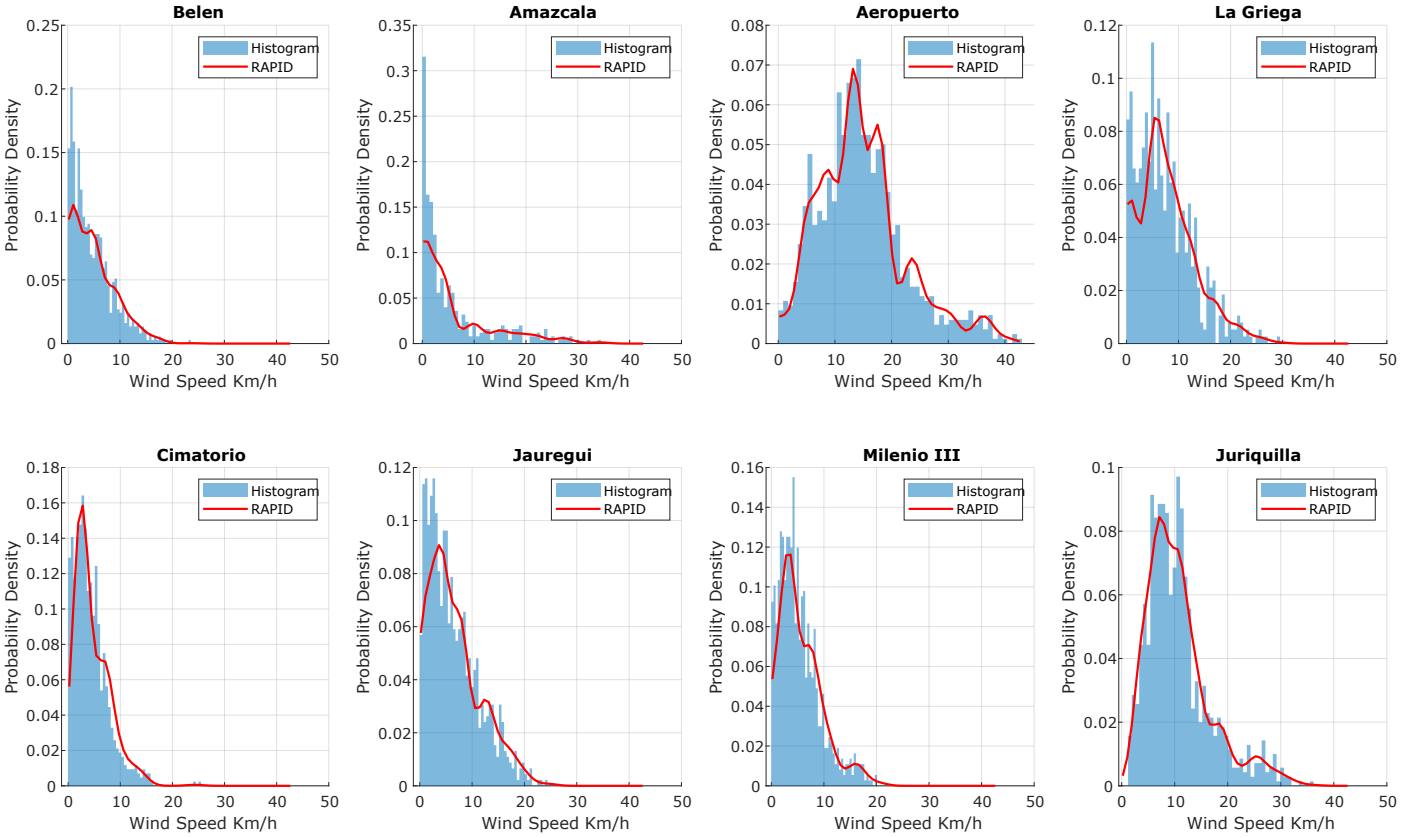


Figure 5. Density estimation using the proposed model.

To evaluate the quality of the fit, point-by-point residuals were defined on the G grid:

$$\epsilon(g_j) = f_{\text{emp}}(g_j) - \widehat{f}(g_j), \qquad g_j \in G. \tag{20}$$

Figure 6 shows the distribution of the residuals, with a strong concentration near zero.

Table 5 summarizes the statistical moments of the residuals. At all sites, the mean remains close to zero, suggesting no systematic bias. On the other hand, sites such as Amazcala exhibit high asymmetry and kurtosis, consistent with extreme events and heavy tails. Consequently, residuals are typically not considered strictly normal; instead, they exhibit location-dependent heterogeneity and zero-centeredness.

In general, these results show that, when the underlying stochastic structure is unknown or when there are heavy tails and multimodality linked to local dynamics, estimating the PDF with an adaptive parametric mixture fitted over $X_{\delta,s}$ is a robust option.

Table 3 reports the final number of components K_s generated by RAPID at each site. This value summarizes the structural complexity required to represent the wind density in the dominant regime captured by $X_{\delta,s}$. It can be seen that K_s ranges from 5 to 10 components, reflecting local differences in multimodality, asymmetry, and heavy-tailed distributions. Sites such as Juriquilla and Aeropuerto, for instance, have $K_s = 10$, which is associated with more intricate structures. Sites like Cimatario, on the other hand, alter the mixture with $K_s = 5$, which is consistent with a more compact probabilistic structure. This behavior is direct evidence of the dynamic growth mechanism controlled by λ in RAPID, without imposing a priori a fixed number of components.

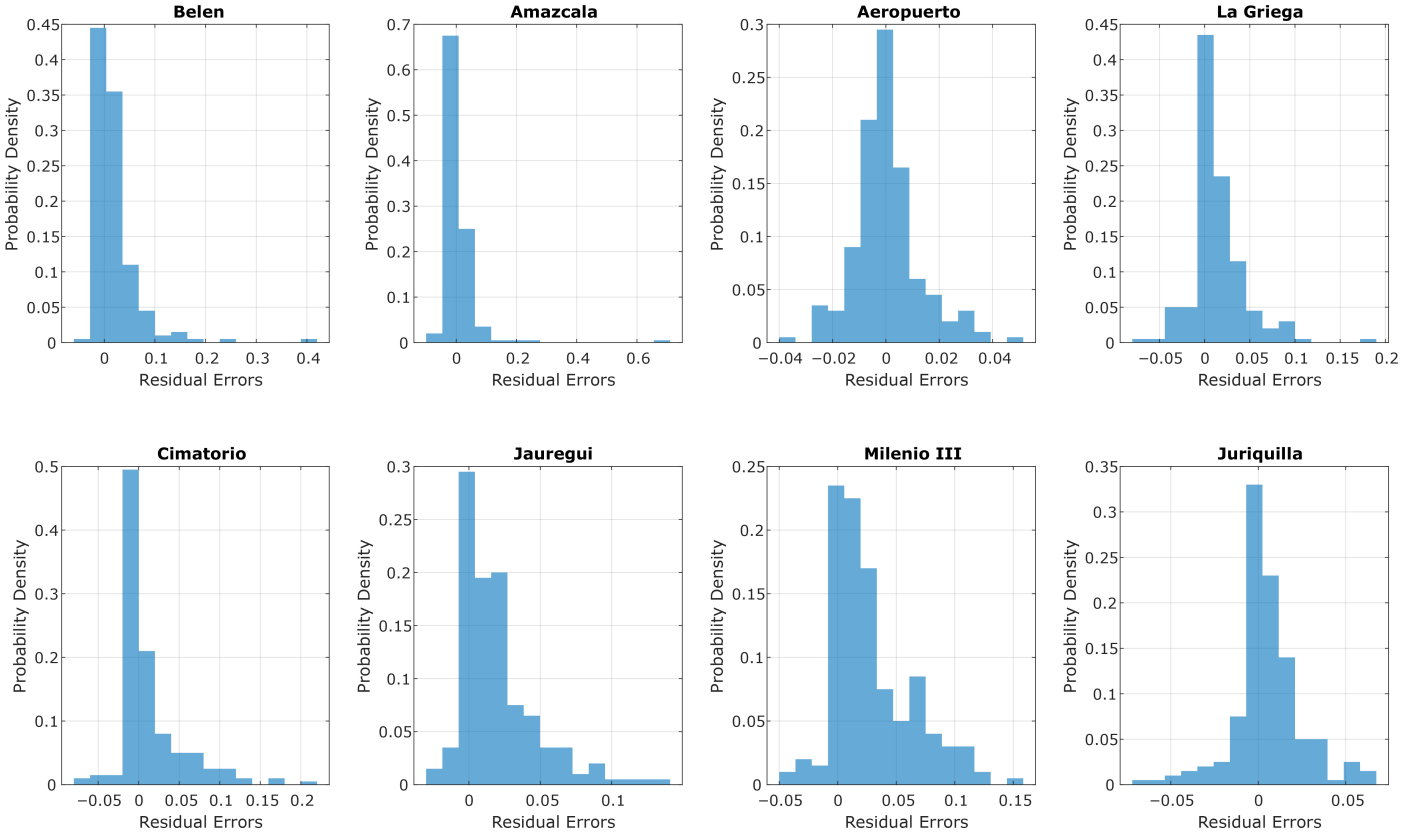


Figure 6. Distribution of residual errors.

Table 5. Statistical moments of residual errors.

Sites	μ	σ^2	γ_1	γ_2
Belén	0.021661	0.000474	1.118	3.783
Amazcala	0.009762	0.000983	4.908	30.352
Aeropuerto	0.000404	0.000033	1.265	6.664
La Griega	0.012405	0.000203	0.645	2.521
Cimatario	0.016718	0.000604	1.045	3.469
Jáuregui	0.019236	0.000200	0.259	1.992
Milenio III	0.028155	0.000711	0.923	3.195
Juriquilla	0.005479	0.000077	0.565	3.911

4.3.3. Sensitivity of RAPID to the Membership Threshold λ

Since the threshold λ directly controls the creation of new components in RAPID, a sensitivity analysis was conducted to examine how the estimator responds to changes in this hyperparameter. In this experiment, the initialization variance was kept fixed at $\sigma_0^2 = 1$, and only λ was varied. The analysis was carried out for three representative sites: Amazcala, Aeropuerto, and La Griega. These sites were selected because they exhibit distinct stochastic behaviors, including heavy tails, extreme events, and comparatively more structured wind regimes.

For each site, RAPID was fitted over the corresponding series $X_{\delta,s}$ using the same preprocessing and evaluation protocol described above. The threshold λ was varied over the grid

$$\lambda \in \{2.0,\ 2.25,\ 2.5,\ 2.75,\ 3.0,\ 3.25,\ 3.5,\ 3.75,\ 4.0,\ 4.25,\ 4.5\}.$$

For every value of λ, the final number of components K_s, the RMSE, and the MISE were recorded. Among the RAPID hyperparameters, λ was prioritized because it is the parameter most directly associated with structural growth, whereas σ_0^2 mainly affects the initialization scale of newly created components.

Figure 7 shows the response of RAPID to changes in λ for the three selected sites. In general, smaller values of λ promote more aggressive component creation, which increases K_s, whereas larger values lead to more compact mixtures.

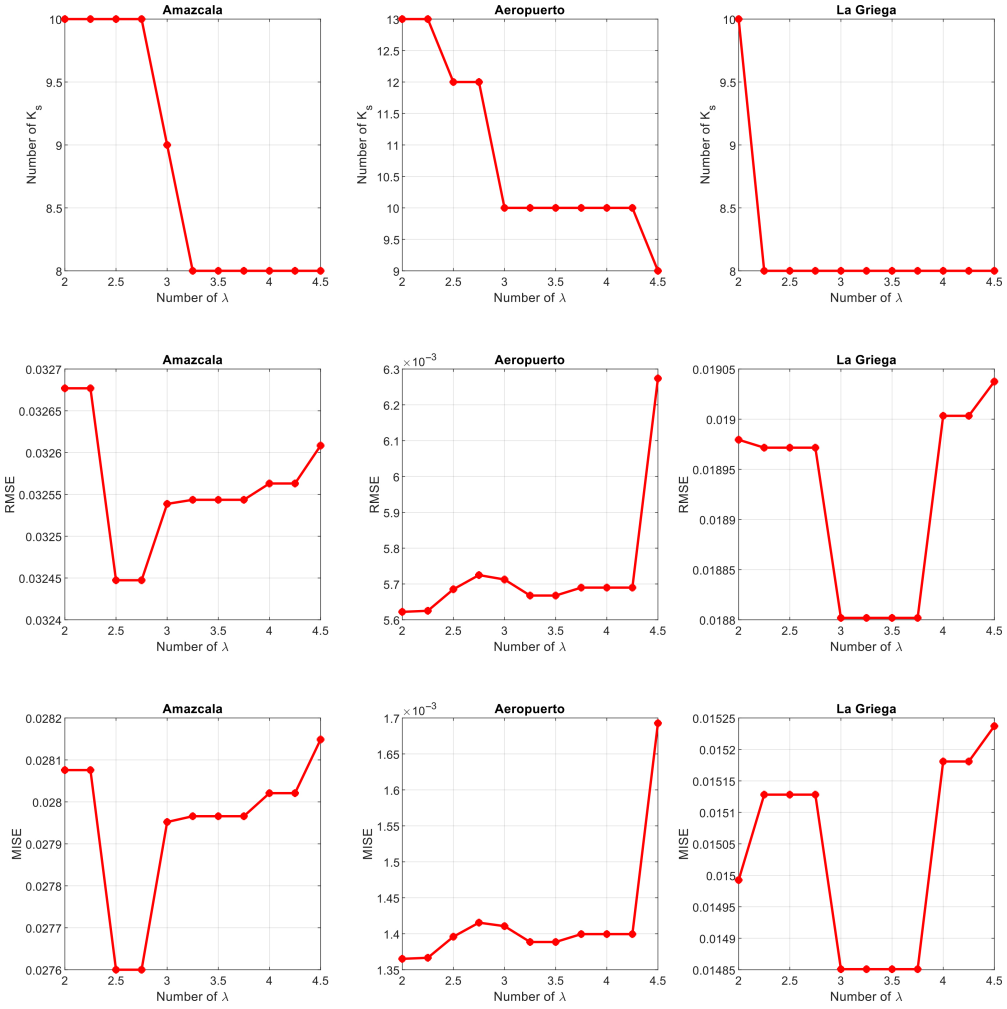


Figure 7. Sensitivity of RAPID to the membership threshold λ for Amazcala, Aeropuerto, and La Griega. The figure shows the variation in the final number of components K_s, RMSE, and MISE under fixed $\sigma_0^2 = 1$ and the same input series $X_{\delta,s}$ for each site.

Table 6 summarizes the numerical results obtained for the three sites under the tested values of λ.

Table 6. Sensitivity analysis of RAPID with respect to λ under fixed $\sigma_0^2 = 1$.

λ	Amazcala			Aeropuerto			La Griega		
	K_s	RMSE	MISE	K_s	RMSE	MISE	K_s	RMSE	MISE
2.00	10	0.032677	0.028076	13	0.005622	0.001365	10	0.018979	0.014992
2.25	10	0.032677	0.028076	13	0.005625	0.001366	8	0.018972	0.015128
2.50	10	0.032447	0.027600	12	0.005685	0.001395	8	0.018972	0.015128
2.75	10	0.032447	0.027600	12	0.005724	0.001415	8	0.018972	0.015128
3.00	9	0.032539	0.027952	10	0.005712	0.001410	8	0.018802	0.014851
3.25	8	0.032544	0.027966	10	0.005667	0.001388	8	0.018802	0.014851
3.50	8	0.032544	0.027966	10	0.005667	0.001388	8	0.018802	0.014851
3.75	8	0.032544	0.027966	10	0.005689	0.001399	8	0.018802	0.014851
4.00	8	0.032563	0.028021	10	0.005689	0.001399	8	0.019003	0.015181
4.25	8	0.032563	0.028021	10	0.005689	0.001399	8	0.019003	0.015181
4.50	8	0.032608	0.028149	9	0.006273	0.001692	8	0.019037	0.015237

The sensitivity analysis reveals a clear trade-off between structural complexity and fitting quality. For Amazcala, smaller values of λ produced a more flexible mixture, with $K_s = 10$ for $\lambda \in [2.0, 2.75]$, and this configuration also yielded the best fitting scores, with a minimum RMSE of 0.032447 and a minimum MISE of 0.027600 at $\lambda \in \{2.5, 2.75\}$. As λ increased, the number of components decreased from 10 to 9 and then to 8, while both RMSE and MISE became slightly worse. This behavior suggests that, for a heavier-tailed site such as Amazcala, a smaller threshold can improve the representation of local irregularities, although the gains remain moderate.

For Aeropuerto, the response was smoother over most of the grid, although the largest threshold produced a visible deterioration. The number of components decreased from $K_s = 13$ at $\lambda \in \{2.0, 2.25\}$ to $K_s = 12$ at $\lambda \in \{2.5, 2.75\}$, then to $K_s = 10$ over most of the intermediate range, and finally to $K_s = 9$ at $\lambda = 4.5$. The best fit was obtained at $\lambda = 2.0$ (RMSE = 0.005622, MISE = 0.001365), but the values observed around the baseline $\lambda = 3$ remained very close (RMSE = 0.005712, MISE = 0.001410) while using fewer components. This indicates that $\lambda = 3$ still provides a favorable complexity–accuracy compromise for this site.

For La Griega, the response was more stable and displayed a broad intermediate optimum. Except for $\lambda = 2.0$, where $K_s = 10$, the estimator remained at $K_s = 8$ across the explored range. The best performance was attained for $\lambda \in [3.0, 3.75]$, where both RMSE and MISE reached their minimum values (RMSE = 0.018802, MISE = 0.014851). In this case, the baseline configuration $\lambda = 3$ lies exactly on the best-performing plateau, indicating that the original choice already provides an adequate balance between compactness and fidelity.

Across the three sites, the overall trend indicates that K_s decreases monotonically or piecewise monotonically as λ increases, confirming that λ acts as a structural growth-control parameter in RAPID. In contrast, RMSE and MISE vary within relatively narrow ranges over the intermediate region of the grid, which suggests that the estimator is reasonably stable to moderate changes in λ. The main deviations appear at the extremes: smaller values increase flexibility through additional components, whereas excessively large values may oversimplify the mixture and degrade accuracy. These results support the interpretation of λ as a growth-control parameter in RAPID: lower values increase flexibility by creating more components, whereas larger values favor simpler mixtures at the possible cost of reduced fidelity in irregular tails or secondary modes.

From a methodological perspective, this analysis supports the use of a fixed global value $\lambda = 3$ as a homogeneous comparison protocol, since it remains at or near the best compromise across the three representative sites. At the same time, the results also show that site-specific tuning may yield modest improvements, particularly in more demanding distributions such as Amazcala or in cases where a slightly richer mixture can be afforded. Therefore, adaptive or data-driven selection of λ remains an immediate direction for future work.

4.3.4. Comparison with State-of-the-Art Alternative Methods

To assess RAPID's performance, it was directly compared with KDE and MoG, which are widely used nonparametric and parametric methods, respectively. In all situations, the three methods were calibrated using the same set $X_{\delta,s}$ to ensure a uniform comparison.

In KDE, a Gaussian kernel with bandwidth selected using Silverman's rule [34] was used. In MoG, the model was fitted using EM [35] and the number of components was selected by BIC evaluating $K \in \{1, \ldots, K_{\max}\}$ with $K_{\max} = 12$. In contrast, RAPID dynamically adjusts the number of components according to the local data structure.

Figure 8 shows a qualitative comparison between fits. KDE tends to smooth the density and may underestimate secondary modes. MoG is sensitive to the number of components determined by the selection criterion, as well as to fit stability in the presence of heterogeneity. On the other hand, RAPID distributes the components to match the local structure of $X_{\delta,s}$, preserving local modes and important asymmetries.

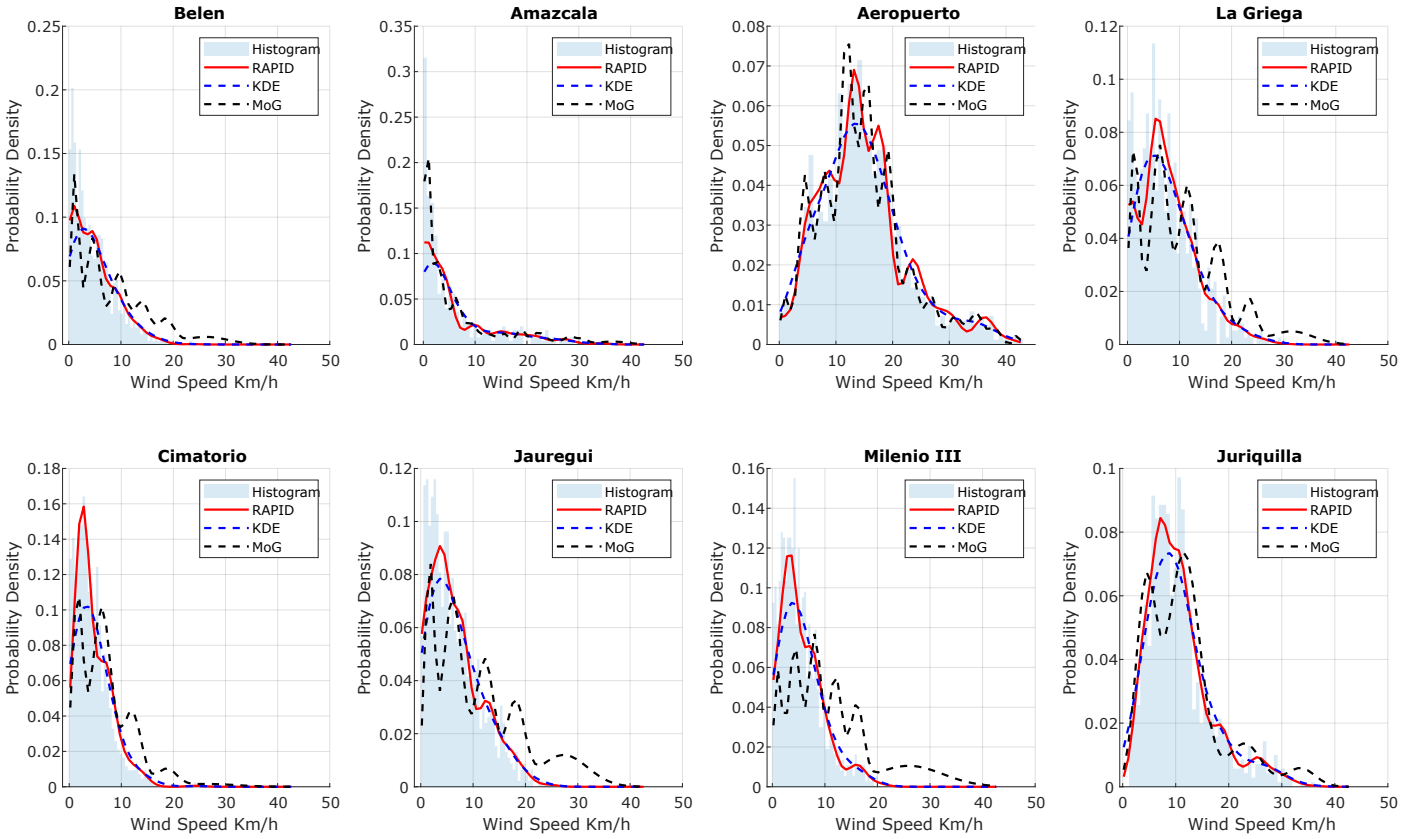


Figure 8. Qualitative comparison of density adjustments.

Table 7 presents the quantitative comparison with RMSE with respect to f_{emp}. In most locations, RAPID yields values comparable to or lower than those of KDE and MoG. Nevertheless, in situations with heavy tails and high kurtosis, such as Amazcala, lower RMSE may favor estimators that approximate the average density shape, even when

relevant aspects of the local stochastic structure are not represented equally well. Therefore, the results are interpreted as evidence of local adaptation and the estimator's structural consistency, rather than as an exclusive optimization of a global metric.

Table 7. Quantitative comparison between RAPID and state-of-the-art methods (RMSE with respect to f_{emp}), numbers in bold represent minima RMSE.

Sites	RAPID	KDE	MoG
Belén	**0.030564**	0.033581	0.031644
Amazcala	0.032539	0.037264	**0.023857**
Aeropuerto	**0.005712**	0.005709	0.007656
La Griega	**0.018802**	0.019337	0.018811
Cimatario	**0.029530**	0.029534	0.031203
Jáuregui	**0.023801**	0.024636	0.026583
Milenio III	0.038597	**0.038142**	0.043519
Juriquilla	**0.010316**	0.011145	0.013504

Table 8 reports MISE (Equation (19)), which quantifies the integrated quadratic error between densities along the support. In general, RAPID achieves the lowest MISE at 5 of 8 sites (Belén, Aeropuerto, Cimatario, Jáuregui, and Juriquilla), supporting a globally competitive fit under the common evaluation protocol. At the Aeropuerto site, KDE and RAPID are practically identical in integrated terms. In contrast, at Amazcala, La Griega, and Milenio III, MoG attains the lowest MISE, which is consistent with more demanding distributional situations involving heavy tails or multimodality, where a batch-optimized mixture with K selected by BIC can reduce integrated error more effectively. In the particular case of Milenio III, RAPID is outperformed by KDE in both RMSE and MISE, and by MoG in MISE, which suggests that recursive local adaptation does not always preserve either pointwise accuracy or global integrated fidelity under the same local regime.

Table 8. Quantitative comparison between RAPID and state-of-the-art methods (MISE with respect to f_{emp}). Numbers in bold represent minima MISE.

Sites	RAPID	KDE	MoG
Belén	**0.039068**	0.04573	0.039632
Amazcala	0.027952	0.036091	**0.016608**
Aeropuerto	**0.001410**	0.001411	0.002534
La Griega	0.014851	0.015344	**0.014307**
Cimatario	**0.035427**	0.036202	0.039056
Jáuregui	**0.024507**	0.026240	0.030077
Milenio III	0.063789	0.062359	**0.059248**
Juriquilla	**0.004535**	0.005368	0.007843

In addition to the asymptotic complexity discussed in Section 3.5, the observed complexity is reported using average execution times per site (Table 9). The measurements were performed in MATLAB R2019b (Windows 64-bit) on a machine with 32 GB of RAM and an Intel(R) Core(TM) Ultra 7155H (1.40 GHz) processor. In all cases, the density estimation methods (RAPID, KDE, and MoG) were adjusted to the same input $X_{\delta,s}$ and non-comparable tasks (file reading, figure generation, and disk writing) were excluded. Times were calculated with $R = 30$ repetitions per site and are reported as the mean μ plus/minus one standard deviation σ in seconds (Table 9).

Table 9. Average execution time and uncertainty by location.

Sites	RAPID (s)	KDE (s)	MoG (s)
Belén	0.000498 ± 0.000075	0.001144 ± 0.000092	0.005801 ± 0.000658
Amazcala	0.000312 ± 0.000021	0.001381 ± 0.000092	0.006886 ± 0.000465
Aeropuerto	0.000648 ± 0.000031	0.001469 ± 0.000072	0.006113 ± 0.000274
La Griega	0.000431 ± 0.000023	0.001216 ± 0.000073	0.005898 ± 0.000503
Cimatario	0.000490 ± 0.000019	0.001105 ± 0.000060	0.005998 ± 0.000359
Jáuregui	0.000608 ± 0.000039	0.001278 ± 0.000080	0.005908 ± 0.000435
Milenio III	0.001266 ± 0.000625	0.002558 ± 0.001227	0.011547 ± 0.005496
Juriquilla	0.001841 ± 0.000618	0.004099 ± 0.000766	0.014961 ± 0.002775

Since ISR is run only once per site to estimate δ and construct $X_{\delta,s}$, the cost reported in Table 9 corresponds to the density fitting (comparative phase) and not to the common preprocessing. To facilitate independent comparison of units, the relative time normalized with respect to RAPID is also reported, defined as $t_{\text{method}}/t_{\text{RAPID}}$ (Table 10). Under this normalization, KDE typically requires between ∼2–4× the time of RAPID, while MoG requires between ∼8–22× as much time, confirming the computational advantage of RAPID's incremental scheme when all methods operate on the same subseries $X_{\delta,s}$.

Table 10. Relative execution time per site, normalized with respect to RAPID (t_{RAPID}).

Sites	$t_{\text{KDE}}/t_{\text{RAPID}}$	$t_{\text{MoG}}/t_{\text{RAPID}}$
Belén	2.30×	11.65×
Amazcala	4.43×	22.07×
Aeropuerto	2.27×	9.43×
La Griega	2.82×	13.68×
Cimatario	2.26×	12.24×
Jáuregui	2.10×	9.72×
Milenio III	2.02×	9.12×
Juriquilla	2.23×	8.13×

5. Discussion

A central result of this study is the marked spatial variability in the dominant wind regime across the MAQ sites, as shown in Table 4. This finding is consistent with the literature on environmental time series, where long records often fail to satisfy global stationarity assumptions but may still admit locally stationary approximations over shorter intervals [11,12,14]. In the present case, the differences observed among stations indicate that wind dynamics in the MAQ should not be represented through a single homogeneous probabilistic structure, since local exposure, relief, and urban conditions affect both temporal behavior and distributional shape.

From this perspective, ISR is not only a preprocessing stage but also an interpretative component of the framework. By identifying a dominant period for each site and validating the induced subsampling through the ADF test, ISR establishes a connection between temporal structure and subsequent density estimation. This is relevant because, when temporal heterogeneity is ignored, a global fit may combine observations from distinct local regimes and distort the resulting PDF. Therefore, the use of $X_{\delta,s}$ provides a more coherent basis for stochastic modeling and is aligned with the local-stationarity rationale reported in the literature [12,14].

The residual analysis supports this interpretation. Although several stations exhibit residuals centered near zero, others show asymmetry and heavy tails, as in the case of Amazcala (Table 5), suggesting that local wind behavior is not only heterogeneous in time but also structurally complex in distributional terms. Under such conditions, RAPID offers

a practical advantage because it does not require the number of mixture components to be fixed in advance. Instead, the model adapts its structure through recursive updating and dynamic component creation, yielding $K_s \in [5, 10]$ across sites (Table 3). This behavior is consistent with the broader literature on adaptive and mixture-based modeling, where recursive updating and the ability to represent multimodality are especially valuable in complex stochastic environments [16,17,30].

Under the homogeneous comparison protocol defined in Section 4.2, RAPID achieved the lowest RMSE at 6 out of 8 sites (Table 7) and the lowest MISE at 5 out of 8 sites (Table 8). These results suggest that the main strength of RAPID is not the optimization of a single metric in isolation, but rather its capacity to maintain a favorable balance among local adaptability, structural interpretability, and fit stability. In contrast, KDE depends strongly on bandwidth selection, whereas MoG requires model selection and iterative EM fitting over a predefined range of components. Within this comparison, RAPID preserves the interpretability of a parametric mixture while avoiding the need to predefine a fixed value of K.

The sites at which MoG attained lower MISE values, particularly Amazcala, La Griega, and Milenio III, are also informative. These sites appear to involve more demanding distributional configurations, where the integrated error is more sensitive to tail behavior, multimodality, or the allocation of mixture components. Rather than contradicting the proposed method, these results help delimit the conditions under which further refinement of RAPID may be beneficial. In particular, they suggest that an automatic adjustment of λ and a more systematic optimization of the initialization parameters $(\phi_0, x_1, \sigma_0^2)$ could improve integrated fidelity without sacrificing the adaptive behavior already observed. These cases also suggest that the sequential nature of RAPID may entail a trade-off between local adaptability and global integration fidelity.

This pattern also makes explicit what may be termed the cost of recursivity in RAPID. Because the estimator is built sequentially, each update is decided from the current component configuration and the incoming observation, which favors adaptive growth and low computational cost. However, once new components are created, RAPID does not perform a full joint re-optimization of all means, variances, and weights over the complete support. As a result, the method may lose global integration fidelity in comparison with batch-optimized mixtures, particularly in tails or low-density regions between components. In some demanding sites, such as Milenio III, this limitation may also affect pointwise fit, as reflected by the fact that RAPID is outperformed by KDE in both RMSE and MISE. By contrast, the batch EM–BIC strategy used in MoG optimizes the mixture globally and can therefore redistribute component structure in a way that better reduces integrated error across the support. In this sense, the lower MISE values attained by MoG at some sites, together with the Milenio III behavior relative to KDE, should be interpreted as evidence of the trade-off between recursive local adaptability and global mixture optimization.

Another relevant advantage of RAPID is its computational profile. As shown in Table 9, RAPID was consistently more efficient than KDE and MoG under the same experimental conditions. This difference is methodologically meaningful because the comparison used the same input $X_{\delta,s}$, the same evaluation grid, and the same discrepancy metrics. The lower cost of RAPID follows from its sequential assignment-and-update mechanism, whereas KDE requires global smoothing and MoG involves repeated EM fitting together with BIC-based model selection. This feature makes RAPID especially attractive for operational or continuously updated settings, where recalibration must be carried out at low computational cost.

To summarize these methodological differences under the common evaluation protocol, Table 11 compares RAPID with KDE and MoG in terms of estimator type, structural adaptability, interpretability, sensitivity to hyperparameters, and computational profile.

Table 11. Comparative synthesis between RAPID and reference density-estimation techniques under the common evaluation protocol.

Criterion	RAPID	KDE	MoG
Nature of the estimator	Adaptive parametric mixture	Non-parametric kernel estimator	Parametric mixture
Comparison input	$X_{\delta,s}$	$X_{\delta,s}$	$X_{\delta,s}$
Structure/components	Dynamic (K_s varies by site)	Not applicable (h controls smoothing)	Fixed after BIC-based selection ($K \leq K_{\max}$)
Sensitivity to hyperparameters	Medium ($\lambda, \sigma_0^2, \rho$)	High (h)	High (K selection and EM convergence)
Ability to represent heavy tails/multimodality	Adaptive (depends on λ and sequential growth)	Sensitive to bandwidth; may oversmooth multimodality	Depends on K and EM convergence
Interpretability	High (components, weights, local complexity)	Low to medium	Medium
Computational profile	Low (typical $\mathcal{O}(M)$; see Section 3.5 and Table 9)	Medium (typical $\mathcal{O}(Mm)$; with fixed m, effective $\mathcal{O}(M)$)	High (EM + BIC: $\mathcal{O}(IKM)$ plus evaluation over $K \in [1, K_{\max}]$; see Table 9)

Overall, the discussion supports two main conclusions. First, the MAQ case study confirms that local stochastic structure matters when modeling urban wind behavior. Second, RAPID provides a competitive alternative to standard reference methods by combining adaptive structure, recursive estimation, interpretability, and reduced computational cost. These properties position the method as a useful site-specific modeling strategy for non-stationary environmental signals.

In Symmetry Perspective

Within this work, the notion of symmetry is restricted to an approximate local invariance under time translation. Let $X_s(t)$ denote the wind-speed process at site s. For a local regime $\mathcal{R}_s$ identified by ISR, this invariance is understood in the weak stochastic sense: for admissible shifts τ within $\mathcal{R}_s$,

$$\mathbb{E}[X_s(t)] \approx \mathbb{E}[X_s(t+\tau)], \qquad \mathrm{Var}[X_s(t)] \approx \mathrm{Var}[X_s(t+\tau)],$$

and

$$\mathrm{Cov}(X_s(t), X_s(t+h)) \approx \mathrm{Cov}(X_s(t+\tau), X_s(t+\tau+h)),$$

for relevant lags h inside the same regime. Under this interpretation, the symmetry considered here is not geometric or flow-based, but temporal and statistical, and corresponds to the local quasi-stationarity assumption used by SICABI.

From this perspective, ISR searches for a time scale δ_s at which the process can be represented through approximately invariant first- and second-order statistics. Thus, δ_s acts as a descriptor of the local regime and establishes the temporal scale at which a weak form of translational invariance is considered plausible.

Departures from this approximate local invariance are reflected in residual asymmetry, heavy tails, multimodality, and cross-site heterogeneity. In this sense, RAPID does not

model symmetry breaking in a formal group-theoretic way; rather, it provides an adaptive mechanism to represent increasing distributional complexity once a simple locally invariant description is no longer sufficient. Under a fixed estimation protocol, the number of components K_s can therefore be interpreted as a relative complexity indicator: larger values of K_s imply that more adaptive structure is required to represent the local density at site s. A simple surrogate measure of this departure is

$$B_s = K_s - 1,$$

These quantities should be understood as operational indicators of departure from a simple locally invariant regime, not as formal order parameters derived from an explicit symmetry group.

6. Conclusions

This work addressed the estimation of wind-speed PDFs in urban environments under non-stationary and spatially heterogeneous conditions. To this end, SICABI was used as a two-stage framework: ISR identified a dominant local regime through spectral analysis and ADF-based validation, and RAPID estimated the corresponding density through an adaptive parametric mixture. The analysis of eight sites in the Metropolitan Area of Queretaro showed that the dominant period is markedly site-dependent, with values ranging from 37 to 166 days, which confirms that the temporal structure of the signal cannot be represented adequately by a single global regime.

The quantitative results support the proposed approach. RAPID adapted its structural complexity across sites with $K_s \in [5, 10]$, without requiring the number of components to be fixed in advance. Under the common comparison protocol, RAPID achieved the lowest RMSE at six of the eight monitored sites and the lowest MISE at five of the eight sites. In addition, RAPID showed a lower computational cost than KDE and MoG under the same input data and evaluation conditions. These results indicate that the method provides a favorable balance between local adaptability, parametric interpretability, and computational efficiency.

From a modeling perspective, the results show that density estimation based on local stochastic structures constitutes a coherent alternative for non-stationary environmental signals. In contrast to global representations, RAPID adjusts the density according to the regime detected at each site, which makes it better suited to heterogeneous urban wind behavior. Although MoG attained lower MISE values at some sites, the overall results suggest that RAPID is competitive as a site-specific estimator and particularly attractive when adaptive structure and low computational cost are required.

Overall, the MAQ case study supports two main conclusions: first, local stochastic structure matters when modeling urban wind behavior; and second, RAPID provides a practical and interpretable alternative to standard reference methods for density estimation in non-stationary climate signals.

Limitations and Future Work

The present study has several limitations that delimit the current scope of the proposed framework. Firstly, the analysis was conducted in a univariate setting, so the model does not yet represent cross-variable or cross-site dependencies. Secondly, the performance of SICABI depends on the correct identification of the dominant local regime through ISR, which may become less reliable when the signal contains multiple regime transitions, several competing spectral peaks, or more irregular temporal structures. Thirdly, the empirical reference density f_{emp} was approximated through a histogram evaluated on a 50-point regular grid, which may affect the absolute magnitude of RMSE and MISE

depending on the discretization used. Finally, the case study was restricted to eight sites within a specific observation period, which limits the direct generalization of the results to other regions, periods, or meteorological variables.

These limitations suggest several immediate research directions. A first extension is to generalize RAPID toward a multivariate and potentially spatial formulation capable of representing dependencies among variables and stations. A second direction is to incorporate automatic hyperparameter adjustment, particularly for λ and σ_0^2, together with more robust strategies for constructing f_{emp} and the evaluation grid. A third line is to extend ISR so that it can accommodate multiple dominant periodicities, either by selecting several relevant spectral peaks or by combining the current procedure with sliding-window analyses. This extension is especially relevant for those sites where the signal may not be adequately summarized by a single dominant regime. Finally, an applied extension of the framework is its integration into a real-time monitoring environment, where stability, latency, missing-data tolerance, and robustness to acquisition noise can be evaluated under operational conditions.

Author Contributions: Conceptualization, D.C.-E. and H.J.-H.; formal analysis, A.-M.H.-N. and H.J.-H.; methodology, S.R.-R. and J.-L.P.-R.; software, D.C.-E.; supervision, A.-M.H.-N. and H.J.-H.; writing—original draft, H.J.-H. and A.-M.H.-N.; writing—review and editing, D.C.-E. and L.-A.D.-J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: We thank the CIICCTE (Centro de Investigación e Innovación en Ciencias de la Computación y Tecnología Educativa) laboratory at the FIF-UAQ for providing technical and infrastructure support.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ACF	Autocorrelation Function
ADF	Augmented Dickey–Fuller
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
EM	Expectation-Maximization
FFT	Fast Fourier Transform
ISR	Identification of Stationary Regions
KDE	Kernel Density Estimation
KPSS	Kwiatkowski–Phillips–Schmidt–Shin
MAQ	Metropolitan Area of Queretaro
MLE	Maximum Likelihood Estimation
MoG	Mixture of Gaussians
PDF	Probability Density Function
RAPID	Recursive Adaptive Parametric Density Estimation
SICABI	Stochastic Inference of Computationally Adaptive Behavioral Structures

References

1. De la torre Gea, G.; Soto-Zarazúa, G.M.; Guevara-González, R.G.; Rico-García, E. Bayesian networks for defining relationships among climate factors. *Int. J. Phys. Sci.* **2011**, *6*, 4412–4418. [CrossRef]
2. Gasparrini, A.; Guo, Y.; Hashizume, M.; Lavigne, E.; Zanobetti, A.; Schwartz, J.; Tobias, A.; Tong, S.; Rocklöv, J.; Forsberg, B.; et al. Mortality risk attributable to high and low ambient temperature: A multicountry observational study. *Lancet* **2015**, *386*, 369–375. [CrossRef] [PubMed]
3. Bouzembrak, Y.; Marvin, H.J. Impact of drivers of change, including climatic factors, on the occurrence of chemical food safety hazards in fruits and vegetables: A Bayesian Network approach. *Food Control* **2019**, *97*, 67–76. [CrossRef]
4. Kiang, J.E.; Gazoorian, C.; McMillan, H.; Coxon, G.; Le Coz, J.; Westerberg, I.K.; Belleville, A.; Sevrez, D.; Sikorska, A.E.; Petersen-Øverleir, A.; et al. A comparison of methods for streamflow uncertainty estimation. *Water Resour. Res.* **2018**, *54*, 7149–7176. [CrossRef]
5. *Climate Change and Tourism—Responding to Global Challenges*; World Tourism Organization: Madrid, Spain, 2008. [CrossRef]
6. Borunda, M.; Rodríguez-Vázquez, K.; Garduno-Ramirez, R.; de la Cruz-Soto, J.; Antunez-Estrada, J.; Jaramillo, O.A. Long-Term Estimation of Wind Power by Probabilistic Forecast Using Genetic Programming. *Energies* **2020**, *13*, 1885. [CrossRef]
7. Burgos-Peñaloza, J.A.; Lambert-Arista, A.A.; García-Cueto, O.R.; Santillán-Soto, N.; Valenzuela, E.; Flores-Jiménez, D.E. Comparative Analysis of Estimated Small Wind Energy Using Different Probability Distributions in a Desert City in Northwestern Mexico. *Energies* **2024**, *17*, 3323. [CrossRef]
8. Hassol, S.; Torok, S.; Lewis, S.; Luganda, P. (Un)Natural Disasters: Communicating Linkages Between Extreme Events and Climate Change. *WMO Bull.* **2016**, *65*, 2–9.
9. Zenner, E.K.; Teimouri, M. Modeling in Forestry Using Mixture Models Fitted to Grouped and Ungrouped Data. *Forests* **2021**, *12*, 1196. [CrossRef]
10. Trifonova, N.; Karnauskas, M.; Kelble, C. Predicting ecosystem components in the Gulf of Mexico and their responses to climate variability with a dynamic Bayesian network model. *PLoS ONE* **2019**, *14*, e0209257. [CrossRef] [PubMed]
11. Hyndman, R.J.; Athanasopoulos, G. *Forecasting: Principles and Practice*; OTexts: Melbourne, Australia, 2018.
12. Box, G.E.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. *Time Series Analysis: Forecasting and Control*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
13. Xu, M.; Chen, X.; Wu, W.B. Estimation of Dynamic Networks for High-Dimensional Nonstationary Time Series. *Entropy* **2020**, *22*, 55. [CrossRef] [PubMed]
14. Dahlhaus, R. Locally stationary processes. In *Handbook of Statistics*; Elsevier: Amsterdam, The Netherlands, 2012; Volume 30, pp. 351–413.
15. Miao, S.; Li, D.; Gu, Y. Fitting wind speed and wind direction probability distribution using mixture B-spline function. *Sustain. Energy Technol. Assess.* **2023**, *60*, 103513. [CrossRef]
16. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.
17. McLachlan, G.J.; Lee, S.X.; Rathnayake, S.I. Finite mixture models. *Annu. Rev. Stat. Its Appl.* **2019**, *6*, 355–378. [CrossRef]
18. Krivec, T.; Kocijan, J.; Perne, M.; Grašič, B.; Božnar, M.Z.; Mlakar, P. Data-driven method for the improving forecasts of local weather dynamics. *Eng. Appl. Artif. Intell.* **2021**, *105*, 104423. [CrossRef]
19. Li, Q.; Zheng, J.; Yuan, S.; Zhang, L.; Dong, R.; Fu, H. RAV model: Study on urban refined climate environment assessment and ventilation corridors construction. *Build. Environ.* **2024**, *248*, 111080. [CrossRef]
20. Dickey, D.A.; Fuller, W.A. Distribution of the estimators for autoregressive time series with a unit root. *J. Am. Stat. Assoc.* **1979**, *74*, 427–431. [CrossRef] [PubMed]
21. Cooley, J.W.; Tukey, J.W. An algorithm for the machine calculation of complex Fourier series. *Math. Comput.* **1965**, *19*, 297–301. [CrossRef]
22. World Meteorological Organization (WMO). Expert Team on Integrated Urban Services (ET-IUS) and Related Guidance on Integrated Urban Services. Available online: https://wmo.int/site/knowledge-hub/governance/sercom/standing-committee-disaster-risk-reduction-and-early-warning-services/expert-team-integrated-urban-services (accessed on 20 January 2026).
23. World Meteorological Organization (WMO). Guidance on Measuring, Modelling and Monitoring the Urban Climate/Urban Heat Island. Available online: https://urban-climate.org/wp-content/uploads/2024/06/1292_en.pdf (accessed on 20 January 2026).
24. RedCIAQ–Universidad Autónoma de Querétaro (UAQ). Red Meteorológica RedCIAQ-UAQ. Institutional Source/Portal. Available online: http://redciaq.uaq.mx (accessed on 20 January 2026).
25. Sundararajan, D. *The Discrete Fourier Transform: Theory, Algorithms and Applications*; World Scientific: Singapore, 2001.
26. Brigham, E.O. *The Fast Fourier Transform and Its Applications*; Prentice-Hall: Saddle River, NJ, USA, 1988.
27. Kwiatkowski, D.; Phillips, P.C.; Schmidt, P.; Shin, Y. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *J. Econom.* **1992**, *54*, 159–178. [CrossRef]
28. Shumway, R.H.; Stoffer, D.S. *Time Series Analysis and Its Applications*; Springer: Cham, Switzerland, 2025; Volume 5. [CrossRef]

29. Huang, W.; Zhu, X.; Xia, H.; Wu, K. Offshore Wind Energy Assessment with a Clustering Approach to Mixture Model Parameter Estimation. *J. Mar. Sci. Eng.* **2023**, *11*, 2060. [CrossRef]
30. Stauffer, C.; Grimson, W.E.L. Adaptive background mixture models for real-time tracking. In Proceedings of the 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), Fort Collins, CO, USA, 23–25 June 1999; Volume 2, pp. 246–252. [CrossRef]
31. Jiménez-Hernández, H.; González-Barbosa, J.J.; Garcia-Ramírez, T. Detecting abnormal vehicular dynamics at intersections based on an unsupervised learning approach and a stochastic model. *Sensors* **2010**, *10*, 7576–7601. [CrossRef] [PubMed]
32. Seinfeld, J.H.; Pandis, S.N. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*, 3rd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2016.
33. Carpintero Aguilar, I. Calibración del Modelo de Precipitación CRHUDA, para su Acoplamiento al Sistema de Alerta REDCIAQ. Bachelor's Thesis, Universidad Autónoma de Querétaro, Facultad de Ingeniería, Querétaro, México, 2024. Available online: https://ri-ng.uaq.mx/handle/123456789/11263 (accessed on 1 January 2020).
34. Silverman, B.W. *Density Estimation for Statistics and Data Analysis*; Routledge: London, UK, 1998. [CrossRef]
35. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B (Methodol.)* **1977**, *39*, 1–22. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.