Scientific Research Publishing

Atmospheric and Climate Sciences, 2024, 14, 328-353
https://www.scirp.org/journal/acs
ISSN Online: 2160-0422
ISSN Print: 2160-0414

# A Hybrid Model Evaluation Based on PCA Regression Schemes Applied to Seasonal Precipitation Forecast

**Pedro M. González-Jardines, Aleida Rosquete-Estévez, Maibys Sierra-Lorenzo, Arnoldo Bezanilla-Morlot**

Center for Atmospheric Physics, Institute of Meteorology, Havana, Cuba
Email: pedro.met90@gmail.com, maibyssl@gmail.com

**How to cite this paper:** González-Jardines, P.M., Rosquete-Estévez, A., Sierra-Lorenzo, M. and Bezanilla-Morlot, A. (2024) A Hybrid Model Evaluation Based on PCA Regression Schemes Applied to Seasonal Precipitation Forecast. *Atmospheric and Climate Sciences*, **14**, 328-353.
https://doi.org/10.4236/acs.2024.143021

**Received:** May 20, 2024
**Accepted:** July 21, 2024
**Published:** July 24, 2024



## Abstract

Possible changes in the structure and seasonal variability of the subtropical ridge may lead to changes in the rainfall's variability modes over Caribbean region. This generates additional difficulties around water resource planning, therefore, obtaining seasonal prediction models that allow these variations to be characterized in detail, it's a concern, specially for island states. This research proposes the construction of statistical-dynamic models based on PCA regression methods. It is used as predictand the monthly precipitation accumulated, while the predictors (6) are extracted from the ECMWF-SEAS5 ensemble mean forecasts with a lag of one month with respect to the target month. In the construction of the models, two sequential training schemes are evaluated, obtaining that only the shorter preserves the seasonal characteristics of the predictand. The evaluation metrics used, where cell-point and dichotomous methodologies are combined, suggest that the predictors related to sea surface temperatures do not adequately represent the seasonal variability of the predictand, however, others such as the temperature at 850 hPa and the Outgoing Longwave Radiation are represented with a good approximation regardless of the model chosen. In this sense, the models built with the nearest neighbor methodology were the most efficient. Using the individual models with the best results, an ensemble is built that allows improving the individual skill of the models selected as members by correcting the underestimation of precipitation in the dynamic model during the wet season, although problems of overestimation persist for thresholds lower than 50 mm.

## Keywords

Seasonal Forecast, Principal Component Regression, Statistical-Dynamic Models

## 1. Introduction

Currently, the increase in the demands of a growing population, together with extreme manifestations of the climate in large regions, leads to a significant increase in the need for resource planning. In this sense, the water resource becomes more important, which participates in all aspects of human socioeconomic life such as health, food security and energy. These factors, described very briefly, drive the need for long-term forecasts that allow adequate planning of available resources and preparation in advance for the presence of extreme weather conditions.

The target of obtaining seasonal forecasts has numerous difficulties. Different climatic drivers that have been identified and described by several authors ([1]-[7]) generate their own influence on precipitation patterns in a certain region. These modulators manifest themselves at different spatio-temporal scales, implying the quantification of said interaction is an important challenge, which makes long-term modeling difficult.

As an example, the NAO (North Atlantic Oscillation) can be modulated by the tropical forcing associated with the ENSO (El Niño South-Oscillation) and the convection associated with the MJO (Madden-Julia Oscillation), or conditioned by fluctuations in the mean of the western Atlantic jet and the result of these interactions will be decisive in the behavior of precipitation towards both Atlantic coasts [8].

The Central American and Caribbean region, due to its position in the tropical zone, may be affected by the redistribution of typical rainfall, which may be conditioned by changes in seasonal regimes or modifications in the quasi-permanent systems that govern the successive changes of the seasonal conditions in the region.

As an example of this, we can mention the tendency for the Hadley cell to expand [9] towards the poles and with it, arid zones, deserts and the subtropical jets positions also move, with the consequent impact in the patterns related to the spatial distribution and intensity of rainfall in the region.

Several approaches have made it possible to respond to this problem. One of them is the dynamic method, where atmospheric general circulation models (GCMs) are used, either forced on their low boundary with SST anomalies (sea surface temperature) or coupled to an ocean model, being able to generate long-term predictions with a middle range of 6 to 12 months. These systems, not limited by purely linear considerations, can represent the processes that influence seasonal variability in a given region, including those of low frequency or unprecedented climate patterns.

However, it has also been documented that they present disadvantages for their application, such as the high computational cost and the tendency to generate biases in the representation of the mean and standard deviation in variables such as precipitation.

In relation to this tool, researches like ([10]-[13]) have described that the

ECMWF presents difficulties such as the generation of dry biases characterized by deficiencies in the prediction of the propagation of the MJO, as well as cold biases in the equatorial Pacific region, which affects the forecast of the amplitude and intensity of ENSO events. Nor does it adequately represent the troposphere-stratosphere interaction, which has repercussions in the adequate forecast of phenomena associated with the QBO (Quasi-Biennial Oscillation). Despite these deficiencies, it has been used to predict extreme behaviors such as drought with good results.

For example, [14] uses its forecasts to carry out probabilistic monitoring of drought using different indices such as SPI (Standardized Precipitation Index), obtaining that it is necessary to apply an inflation factor to improve the ensemble's standard deviation with respect to the observations. Similar research was carried out in the Latin American region [15], in this case, their conclusions suggest that the model allows representing episodes of moderate drought. However, its performance is not good enough to pro-vide a useful guide in months of large rainfall accumulated.

Despite these results, the use of the statistical approach has predominated in the Latin American region. This methodology uses an empirical relationship between a predictor (variable that is used to make the forecast) and a predictand (variable that is to be predicted). Their main advantage is that they are designed to be consistent with observations and can provide probabilistic and deterministic predictions, in addition to their application requiring low computational resources. However, they assume a stationary climate system and, due to the linearization that characterizes these methods, they tend to have problems adequately representing nonlinear interactions (such as convective processes in the tropics) as well as the standard deviation of the predictands [8].

This approach can be developed in two aspects, one where observations are used to establish functional relationships, often regression, through which predictions can be made, or combined with CGMs in what is known as a hybrid approach.

The use of SST (sea surface temperature) as a predictor to forecast the seasonal variability modes of precipitation (predicting) has been the most commonly used, however, its results show modest skill indices ([16] [17]).

More recent studies [18] find that predictors related to moisture fluxes can more skillfully predict different characteristics of precipitation compared to SSTs. Several authors relate predictors such as vertically integrated moisture transport flux, zonal wind and 850 hPa surface temperature, reduced sea level pressure (slp) and specific humidity lead to results superior to SSTs ([18] [19]).

Several hybrid approaches suggest that the use of ECMWF solutions can be used to train statistical models that allow for improving the skill of the dynamic model in a given region. For example, [20] uses supervised learning methods and neural networks to obtain forecasts of the flow of the Tocantins River. To do this, they apply a bias-correction scheme to the outputs of the dynamic model to

then feed the statistical models. Their results coincide with [21], who finds that the ECMWF is able to represent the rainfall seasonality in the region of interest. Other authors such as [22] and [23] have used other methods such as Bayesian models or wavelet functions to obtain similar results.

[24] uses categorical classification correction models using machine learning techniques based on ECMWF solutions. This research evaluates two methodologies. On the one hand, they use the ECMWF ensemble mean as input data for various machine learning schemes that are evaluated as individual models. The second methodology uses the ensemble members separately to give input to the models, thus generating committee models. Their results suggest that the second approach allows for preserving the physical relationship through individual training of the statistical model with each ensemble member of ECMWF.

At national, work with predictive purposes on seasonal and sub-seasonal scales has been superficial and has been fundamentally oriented to purely statistical approaches. They can be cited ([25] [26]) who suggest the use of a regression model 6 months in advance. On the other hand, [27] try to use a dynamic approach, selecting the WRF (Weather Research and Forecasting) with the objective of predicting precipitation associated with synoptic-scale systems. This last variant, however, has the weakness of being fundamentally based on existing experiences with short- and medium-term forecasting. None of these experiences managed to be fully developed.

Taking this background into account, the objective of this research is to develop and evaluate a hybrid seasonal precipitation prediction model based on principal components analysis (PCA) regression schemes, using the ECMWF ensemble mean as input.

## 2. Materials and Methods

### 2.1. ECMWF-SEAS5 and Predictors Selection

The ECMWF-SEAS5 model's forecast is used to obtain the group of predictors that are subsequently used to feed the statistical models. According to the results of [12], the new European model's version, coded as SEAS5, presents several characteristics in its performance in tropical regions.

According to the authors, the dynamic system tends to generate warm biases in the tropical ocean, particularly in summer, which they suggest is related to the fact that the system tends to produce shallower mixed layers in these regions. The Pacific and Atlantic basins are the most affected relate to the sea surface temperature forecasting. This has an influence on the prediction of large-scale dynamics related to heat and moisture transport in the easterly flow toward the tropical region. Together with the additional cooling produced in the Niño 3.4 region, although to a lesser extent than its predecessor SEAS4, it can lead to rainfall underestimations in the tropical region. The errors in the SST prediction of the Niño region seem to be more notable in the DJF (December-January-

February) trimester.

ECMWF also tends to reduce the gradient between the action centers that give rise to the NAO, which leads to this oscillation not being well represented at the surface. In relation to the dynamic processes that occur in the stratosphere and their relationship with teleconnection phenomena, the model basically manages to predict the phase of the QBO, however, it does not manage to well represent the amplitude and intensity of this oscillation. According to the authors, this is related to the presence of cold biases in the tropical tropopause environment together with a transition towards a warm bias in the high tropopause, which can generate errors in the representation of the mid-latitude jets and the simulation of stronger westerlies above the 40 hPa surface.

This characterization suggests that the model will inadequately represent certain atmospheric dynamical responses with the consequent impact on predicted seasonal rainfall patterns. In this case, the use of alternative methods that improve these predictions seems to be an alternative to obtaining seasonal forecasts with regional applications.

Following this line, the present work proposes to use a series of predictor variables extracted from the ECMWF forecasts, through which statistical models will be built to try to improve the dynamic system forecast. A total of 7 are used, which are derived from the results of [28]. Who use the maximum covariance method (MCA) within a principal components analysis to determine the greatest spatio-temporal associations between several predictors candidates and the monthly rainfall recorded in Cuba.

OLR (Outgoing Longwave Radiation), SLP (Mean sea level pressure), T850 (Temperature at 850 hPa surface), Asst (Tropical Atlantic Sea Surface Temperature) and Csst (Caribbean Sea Surface Temperature) are similar to those used in other research where statistical relationships are used ([17]-[19] [29]). Added to this more traditional group is the Gálvez-Davison Index (GDI).

GDI is a stability index generated to improve the convection forecast in the Caribbean [30]. Research by ([31] [32]) suggests that the use of thermodynamic indices such as CAPE can be useful to explain certain precipitation patterns and their connections with phenomena of seasonal influence on the behavior of the precipitation in the Caribbean region such as the American monsoon. However, the CAPE usually has limitations in the tropical region since its values can be high due to the higher tropopause height in the tropics with respect to other regions, this does not necessarily imply the presence of deep convection regimes in certain circumstances. In contrast, for the Caribbean region, it was found that the GDI can explain more than 50% of the variance of precipitation and represent, with skill values higher than traditional thermodynamic indices, different precipitation regimes [30].

This index considers three physical processes that modulate tropical convection: the simultaneous availability of heat and moisture at the middle and lower troposphere; the stabilizing/destabilizing effects at middle and upper levels caused by ridges and troughs; and dry air entrainment and stabilization related

to inversions (Equation (1))

$$GDI = ECI + MWI + II \tag{1}$$

where the ECI is a stability index, MWI represents the heat content at middle levels and II is an inversion index.

In all cases, the predictors were extracting to forecasts generated by ECMWF with a delay of one month (lag = 1), taking into account the results of [24], who affirm that the forecasts of the model for lag = 0 and lag = 1 lead to the most robust results in relation to rainfall. The years considered for the study correspond to 2020 and 2021; where the selection criterion takes into account differences between the annual cycle recorded in both years. The subregions used for the predictors are listed in (**Figure 1**).

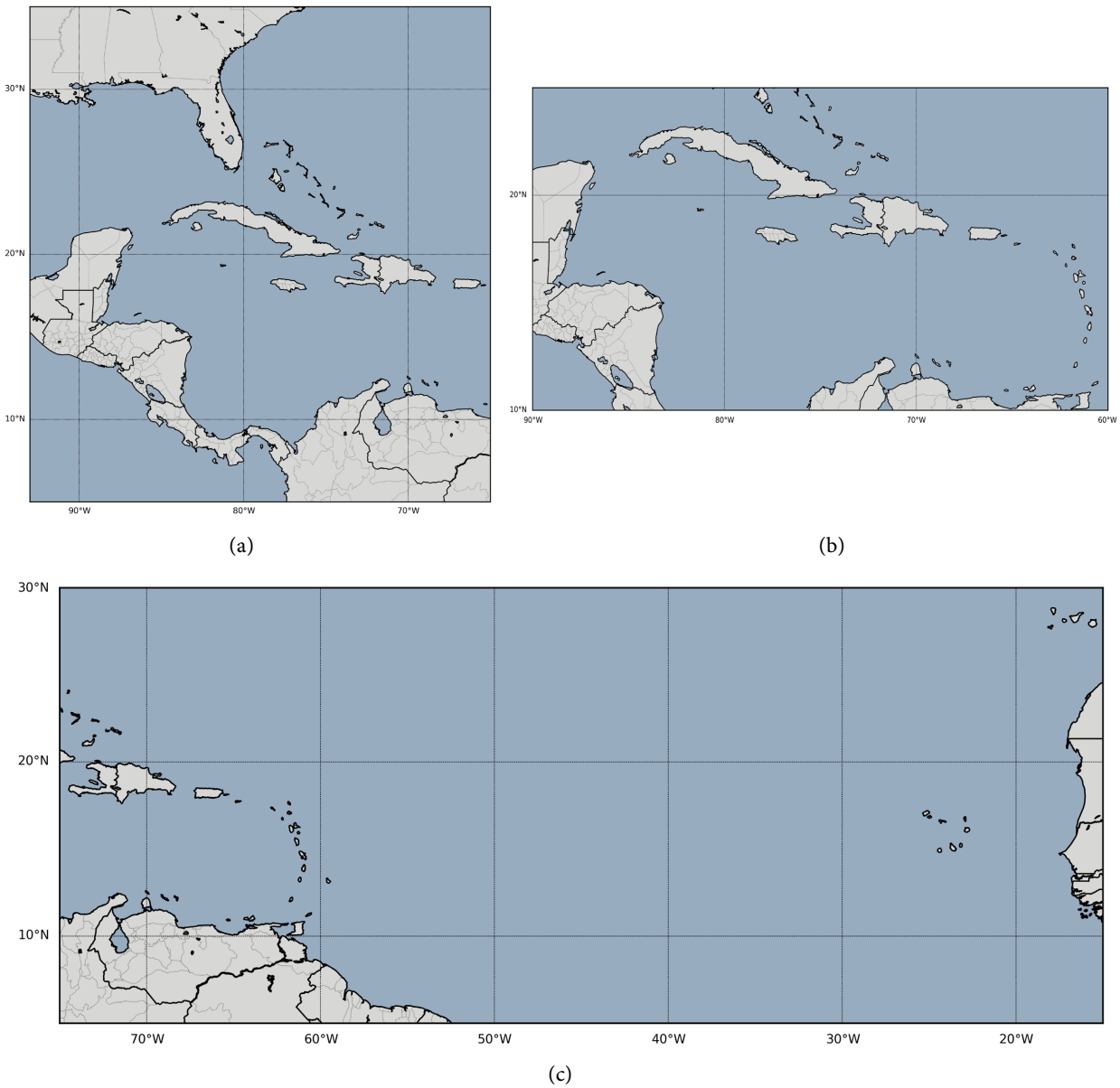



**Figure 1.** Predictors's subregions. (a) OLR, SLP, T850, GDI; (b) Csst; (c) Asst.

## 2.2. Predictand

The monthly accumulated rainfall is selected as the predictand. This selection is due to the fact that a national grid generated from the research of [33] is used (**Figure 2**). These authors use the records of accumulated monthly rainfall recorded by the network of meteorological stations and the pluviometric network (belonging to the National Institute of Hydraulic Resources), which are interpolated to a 4 km (kilometers) grid of spatial resolution. The use of these data allows for the availability of high-resolution observational records from 1980 to the present.

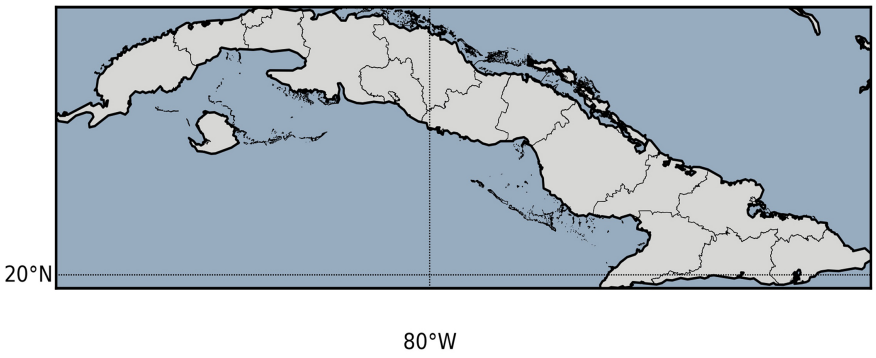



**Figure 2.** Predict and domain, used for the hybrid system forecast too. The values are only present in land areas, while adjacent seas are defined as NaN.

## 2.3. PCA Regression Methods

To generate hybrid forecasts, three PCA regression methods are proposed. Each of them is applied individually to each predictor with the purpose of generating individual models. Prior to using each model both, predictor and predictand, are standardized to subsequently perform a dimensionality reduction using the principal component method (PCA). This step pursues the purpose of representing those patterns with the greatest influence on the variability of precipitation over the study area (**Figure 2**) by eliminating noise and possible multicollinearity errors.

In all cases, a retrospective forecasting strategy is used where two sequential training periods are considered, one of 20 years and another of 6 months prior to the target month, in order to evaluate its possible impact on the hybrid design (**Figure 3**). Since the data is distributed on a monthly scale, this implies that the schemes with the longest training period can be decomposed into a larger number of principal components. Taking these differences into account, an arbitrary restriction is applied to guarantee the greatest possible homogeneity in the designs. This restriction consists of using as many components as necessary to explain more than 90% of the predictor and predictand's variances. In all cases, the constructed models are applied to the corresponding PCA; once the forecasted PCA are obtained, the fields are reconstructed, thus obtaining a deterministic forecast of the predictand.

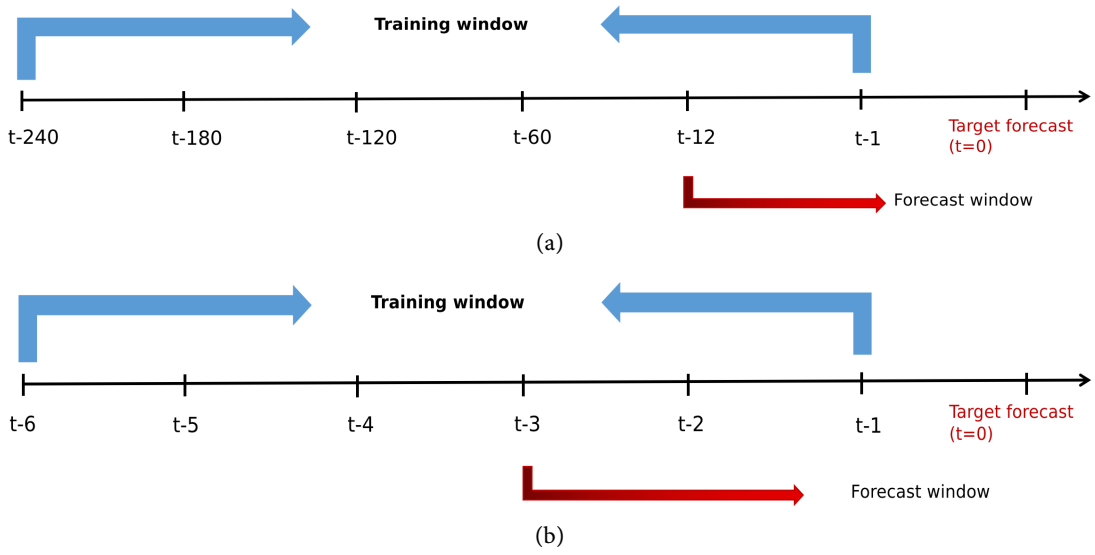



**Figure 3.** Schematization of the selection of training and forecast thresholds; (a) 20 years, (b) 6 months. For all cases using a monthly $t_n$ step.

### 2.3.1. PCA Linear Regression

The method, as it is known, is based on the formulation described in (Equation (2)), where $Y$ represents the predicted PCA; $X$ the predictor's PCA; $\beta_1$ and $\beta_2$ the coefficients and $\varepsilon$ the noise obtained from the model residuals.

$$Y = \beta_1 + \beta_2 X + \varepsilon \tag{2}$$

This linear model is considered a parametric method [34] since it assumes that the relationship between the variables is linear and can be defined by the line parameters, which implies that the changes between the predictor and predictand will be constant. This is an important limitation, since given the non-linear nature of the variables involved, it can lead to biased or erratic predictions that lead to low model skill.

Another disadvantage is that, because it assumes that the predictor-predictand functional relationship is represented by a linear function, it may be deficient in representing extreme or rare events and therefore, this makes it sensitive to the presence of themselves during the training period, which can have negative repercussions on the forecasts.

### 2.3.2. Linear PCA Regression Based on K-Nearest Neighbor

This methodology involves a modification of linear regression in its simplest version, following in this case a non-parametric approach, which implies that it does not establish a prior functional relationship between the predictor and predictand, making it more flexible than the proposal in the section above. However, it offers similar results to linear regression.

The basic nearest neighbor regression obtains a relationship similar to the linear one (Equation (2)), where it calculates the Euclidean distance but limits the number of points (Equation (3a)). To do this, it uses uniform weights: that is, each point in the local neighborhood contributes uniformly to the classification

of a query point. In this case, the points are weighted so that weights are proportional to the inverse of the distance from the query point (Equation (3b)).

$$d = \sqrt{\sum_{i=1}^{k} \left(x_i - y_i\right)^2} \tag{3a}$$

$$f_0 = \frac{1}{k} \sum_{i=1}^{k} y_i \tag{3b}$$

where $d$ represents the distance between the observation and $f_0$ prediction, the weight or weighting of the $k$ selected points in the neighborhood of the point of interest.

The curve building from line segments that are generated in each neighborhood allows the model to not be conditioned by a specific functional relationship as in the previous case. However, the selection of $k$-neighbors can be critical, since large values of $k$ reduce the effect of noise but make the boundaries between classes less clear. Additionally, the method is also sensitive to rare or anomalous events and may not work adequately if the available sample is small [35]. For this work, $k = 5$ is used for 6-month training and $k = 10$ for multidecadal training.

### 2.3.3. PCA Regression Based on Support Vectors

The method is based on the construction of an optimal hyperplane in the form of a decision surface, so that the margin of separation between the two classes in the data is maximized. This means that it will try to fit the best line within a threshold value. The threshold value is represented as the distance between the hyperplane and the limit line trying to satisfy the condition: $-\alpha < y - \beta_2 x + \beta_1 < \alpha$, where $\alpha$ represents the margin or threshold.

This methodology allows solving the regression problem with different approaches, ranging from linear, which is already worked with the methods described above, to more complex ones such as polynomial, sigmoid or RBF (Radial Basis Function). Additionally, the implementation of the method through the Scikit-Learn Python library allows you to create your own kernels. In this case, it is decided to use a nonlinear kernel represented by RBF (Equation (4)).

$$f\left(x; y\right) = \mathrm{e}^{\frac{-\gamma \left\| x_i - y_i \right\|^2}{\sigma}} \tag{4}$$

where the numerator of power represents the Euclidean distance and the denominator is the variance of the hyperplane [35].

When using this function two critical parameters must be considered. One of them is C, which is common for all support vector kernels, which compensates for miss classification of training examples with the simplicity of the decision surface. This implies that small values of this parameter tend to smooth the result, while large values imply the classification of a greater number of examples by selecting more samples as support vectors, in other words, expanding the hyperplane [36].

The second parameter is gamma, which defines how much a single training example influence, and can be interpreted as the inverse of the radius of influ-

ence of the selected examples. As can be seen, the selection of both parameters is critical for the performance of the model; if gamma is too large, the radius of the influence area of the support vectors only includes the support vector itself and no amount of regularization with C will be able to prevent overheating. On the other hand, if gamma is too small, the model is too limited and cannot capture the complexity or shape of the data.

It is usually recommended that both parameters be spaced exponentially between each other, normally a logarithmic spacing between $10^3$ and $10^{-3}$ is usually sufficient [36]. In this study, it was necessary to increase the value of the parameter C to around $10^6$ for all cases, since with lower values very smoothed forecast curves were obtained that did not represent the characteristics of the annual rainfall behavior. However, this led to this methodology being the most computationally expensive compared to the other variants.

#### 2.3.4. Application Limits

It should be taken into account that there are several factors that can limit the application of the methods used in this research and therefore the quality of the results. For example, when building forecast models from principal components, there is a dilemma as to how many components can be included; too many components can lead to overfitting and too few components can lead to not including variability modes that provide precipitation. Another fundamental limitation of this methodology is that principal component analysis and linear regression assume linear relationships between the variables. However, relationships in climate data can be nonlinear, such as tropical rainfall. The proposed design attempts to reduce the impact of these effects on the forecast.

On the other hand, the volume of data is another factor that limits the degree of precision achieved with these methods. The methods in question are highly dependent on the length of the series in which they will be applied. The larger the data sample, the greater the capacity of the model to capture the temporal and spatial behavior of precipitation.

### 2.4. Evaluation Metrics

With the purpose of evaluating the added value of the hybrid design, a group of metrics is selected that will allow establishing the ability of the models built on spatio-temporal scales. The years 2020 and 2021 were selected to apply the evaluation. The selection criterion is based on the difference in relation to the behavior of the annual cycle, being very similar to what is theoretically described for the region in the case of 2020, while the following year resulted in a more discreet behavior in relation to the accumulated monthly recorded, with a downward trend after the May peak.

Bias values (Equation (5)), RSME ((Equation (6)), KGE (Equation (7)) allow it to quantitatively describe the behavior of the built models. Additionally, coefficient of determination values were used to support the selection of the best performing models (Equation (8)). In this case, the first three respond to a cell-point

verification scheme, so the evaluation was complemented with a dichotomous analysis based on the use of SEDS [37] that seeks to more clearly discriminate the thresholds where the proposed models gain/lose skill (Equation (9)).

$$\text{Bias} = \frac{1}{n}\sum_{i=1}^{n}\left(obs_i - fcst_i\right) \tag{5}$$

$$\text{RSME} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\left(obs_i - fcst_i\right)^2} \tag{6}$$

$$\text{KGE} = 1 - \sqrt{\left(r-1\right)^2 + \left(\alpha-1\right)^2 + \left(\beta-1\right)^2} \tag{7}$$

$$R^2 = \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2} \tag{8}$$

In this case, $r$ represents the linear correlation between the forecast and the observation, while $\alpha$ and $\beta$ are the ratio between the standard deviation and the mean respectively calculated between the forecast and the observation; as well as $\sigma_{XY}$ represents the covariance between the prediction and the observed and $\sigma^2$ the respective variances.

$$\text{SEDS} = \frac{\log q - \log H}{\log p + \log H} \tag{9}$$

where $H$ represents the correct detections, while $p = (a + c)/n$ is the relative frequency with which the observed event is detected and $q = (a + b)/n$ is the relative frequency of predicted events; the parameters $a$-$b$-$c$ are obtained from the contingency table for binary events (Table 1); where n is the total number of events.

Table 1. Contingency table model for binary events.

| Forecasted | Observed | |
|---|---|---|
| | *Yes* | *No* |
| *Yes* | *a* | *b* |
| *No* | *c* | *d* |

## 3. Results

### 3.1. Annual Rainfall Cycle

The annual cycle behavior on the island is defined by the presence of a dry season (November-April) characterized by rainfalls fundamentally associated with extratropical systems that invade the tropical zone. At this cause, the predominant influence of dry air masses, related to migratory anticyclones, substantially increases the days without precipitation in these months.

Towards spring, during the transition to summer, the rainfall generally tends to increase, conditioned by the persistence of meridional temperature gradients, the influence of subtropical jet and the westward expansion of the subtropical ridge, which favors the transport of moisture from the ITCZ that is moving northward. Often this combination of factors is also capable of inducing baroc-

linic instability that leads to manifestations of severe weather that provide important records of precipitation in short time periods at this period of year.

Already during the wet season of the year, the first rainfall peak appears, usually between May and June. In this sense, seasonal systems appear that significantly increase rainfall towards the western and central regions of the country, such as the May-June trough. This system, the result of the flow of troughs or short waves from mid-latitudes, is often located between the Mexican ridge and the subtropical ridge between medium and high levels. This configuration is capable of interacting with synoptic-scale systems such as easterly waves, upper lows or the subtropical jet itself and, as a whole, favor the early peak of precipitation.

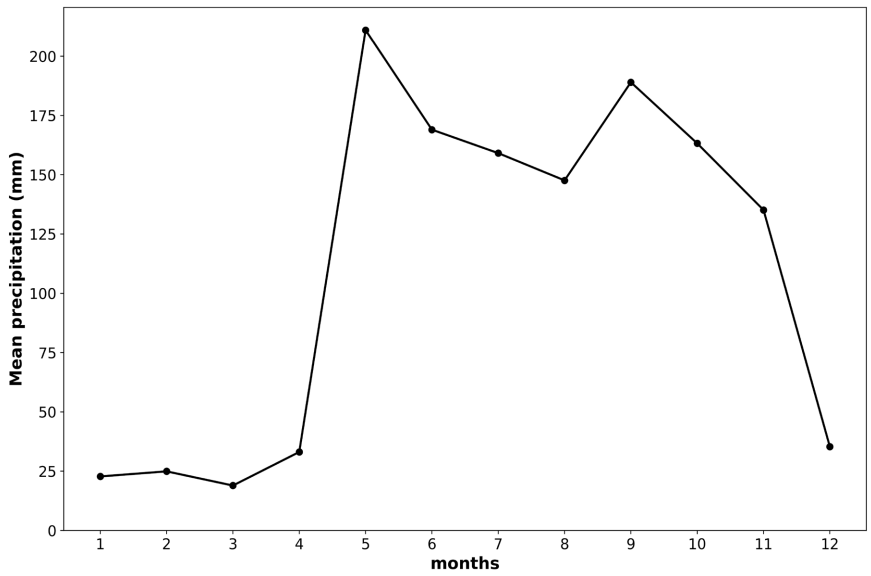
The mid-summer drought process appears as a consequence of the westward expansion of subtropical ridge, conditioning the arrival of dust clouds from the Sahara in the months of July and August where their non-hygroscopic aerosols favor the absence of rainfall. In this case, the ridge completely extended to the west (often to the Gulf of Mexico) tends to cut off moisture transport. In these months, precipitation is usually associated with strong diurnal warming, occasionally stimulated by the passage of tropical waves, which often exhibit discrete representations in their cloud field, or higher lows that usually emerge from the TUTT (Thought Upper Tropical Tropospheric).

Towards wet season ends, the withdrawal of the subtropical ridge, generates a typical minimum sea level pressure in the area, thus allowing greater exchange with the extratropics. At this time the ITCZ has reached its maximum displacement towards the north and begins the southward migration, this is conducent the moisture convergence in the region and also conditioning the peak of the hurricane season in the Atlantic basin, therefore, a new rainfall increase is observed, producing a second maximum.

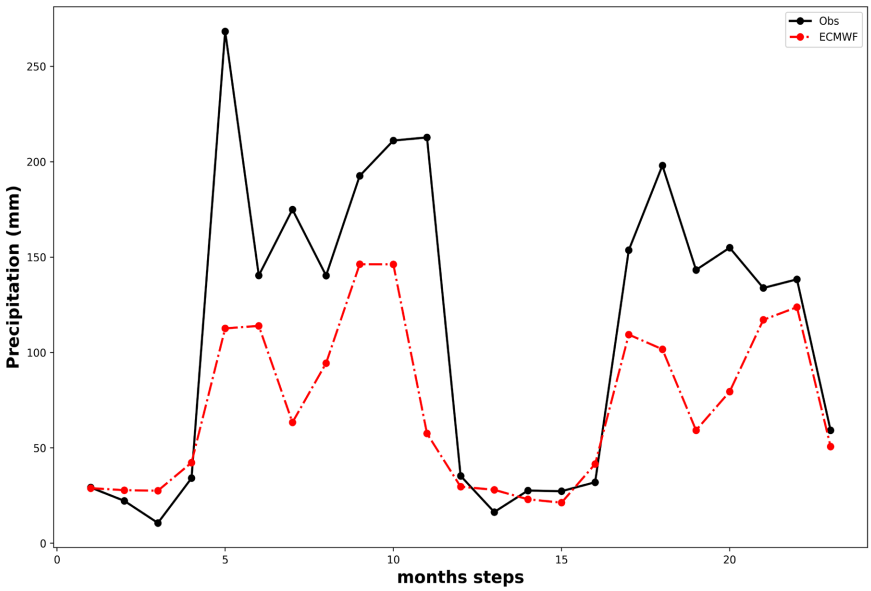
These average characteristics are usually reflected in the average behavior of precipitation in Cuba (**Figure 4(a)**) in the selected years for the research. However, towards 2021 the delay in the contraction of the ridge influences the conditions of subsidence and divergence, which is why a discrete late peak of precipitation occurred on the island, reinforcing the conditions of seasonal drought.

In this sense, coinciding with authors from the region [15], it can be observed that the ECMWF forecasts, estimated from the ensemble mean, underestimate the monthly rainfall accumulated mainly in the wet season. That can be related to deficiencies in adequately simulating the moisture fluxes behavior and divergence advection related to the relative position of subtropical ridge, since it additionally suggests that the late peak is usually more active than the first maximum, which is opposite to the records obtained.

This means that the ECMWF model tends to overestimate moisture transport in the Caribbean during the transition to winter, which may be related to the tendency to produce warm biases in the Atlantic basin in combination with the behavior of ridge contraction [12]. Taken together, these characteristics maintain the ocean-atmosphere feedback and can explain the model's representation of the annual precipitation pattern.




(a)




(b)

**Figure 4.** Rainfall average behavior between 2020 and 2021 (a); annual cycle (black) and ECMWF forecast (red) (b).

The latter can be corroborated in **Figure 5**, where the ECMWF forecast not only inadequately represents the flow, but also exhibits a tendency to overestimate sea level pressure values, this suggests a greater influence of the subsidence associated with the subtropical ridge and lower moisture transport from the ITCZ region. On the other hand, a configuration like the one shown in (**Figure 5(a)**) also reduces the exchange with the extratropics and the instability associated with the establishment of meridional temperature gradients. All of these factors contribute to the underestimation observed in the dynamic model forecast.

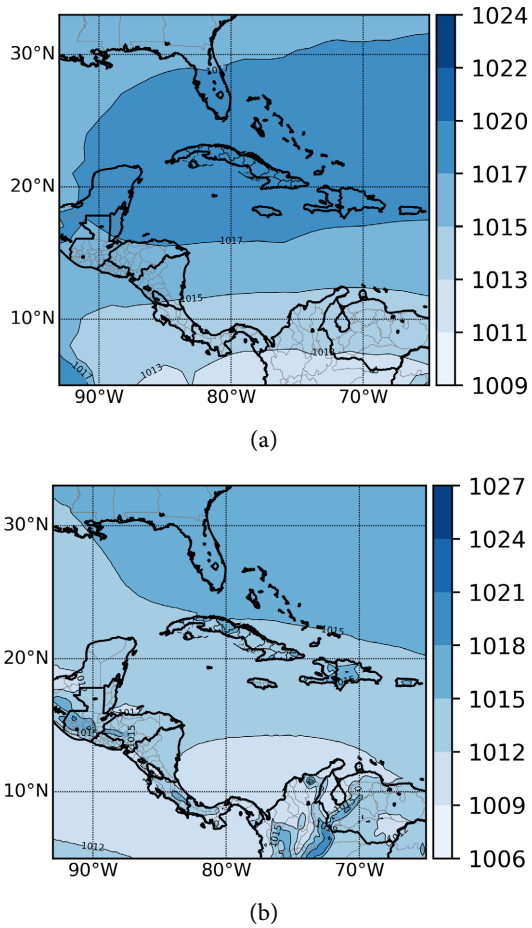



**Figure 5.** Mean sea level pressure behavior in October 2021. (a) ECMWF's ensemble mean; (b) ERA5 reanalysis.

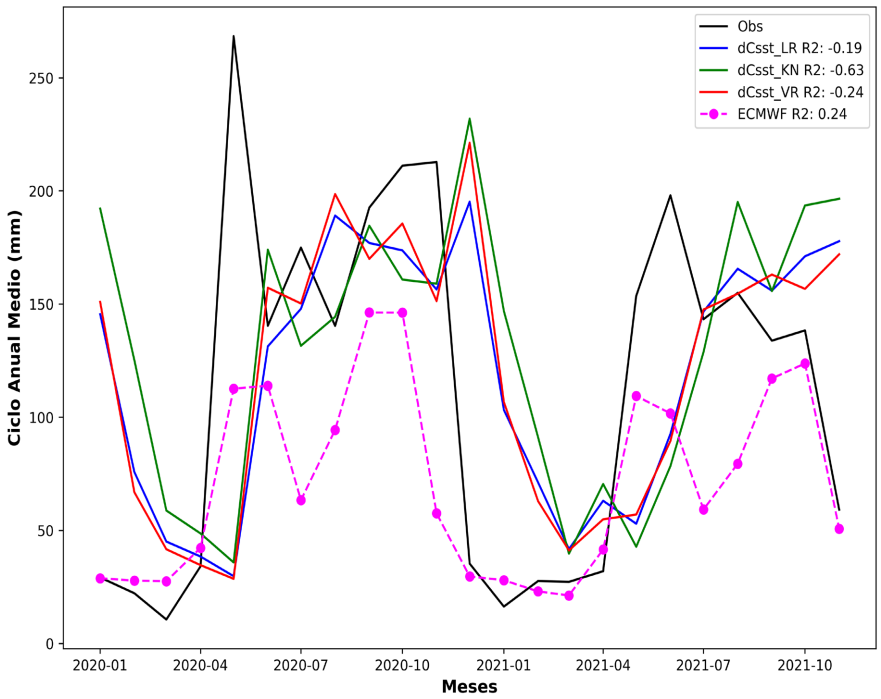
## 3.2. Evaluation of Individual Models

The statistical-dynamic models generated from the proposed methods suggest that long sequential training periods are not suitable for generating monthly forecasts (**Figure 6**). As shown in the graphics, for all the proposed predictors the forecast curves are very smoothed and unrealistic. The determination coefficient values are consistent with these characteristics, generally exhibiting a behavior close to zero or negative.

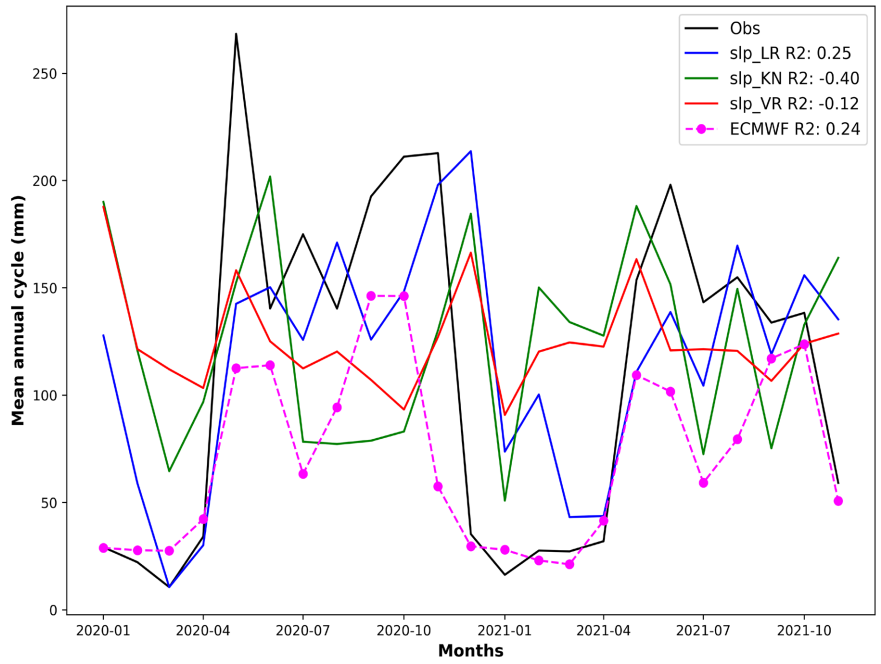
In these cases, only the models built from linear regression showed values of determination coefficient close to those obtained with the ECMWF, as is the case of SLP (**Figure 6(b)**) and OLR. However, this value is purely numerical and has no physical meaning since the predicted curves do not represent any of the seasonal characteristics of rainfall on the island.

On the opposite, the 6-month sequential training generated more coherent results and it was possible to discriminate both methods and predictors based on the individual forecasting ability of each one. The construction of the annual trend for the years selected in this study suggests that the predictors that represent the SSTs have low ability to reproduce the monthly rainfall (**Figure 7**).

These results coincide with [18] who grant greater predictive importance to moisture fluxes compared to SSTs. At this point, the regression coefficients obtained for different PCA regression schemes, capture the seasonal trend of most predictors.

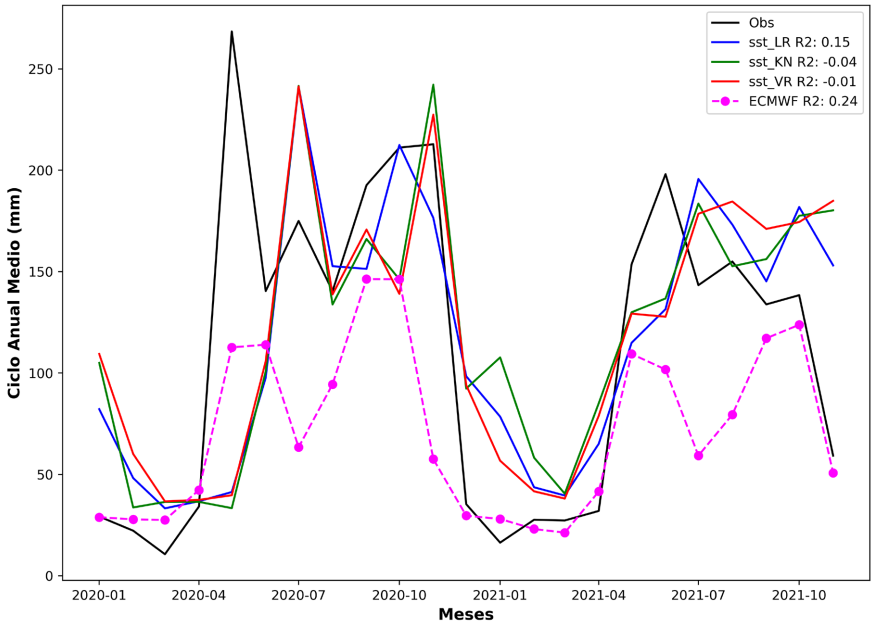



(a)

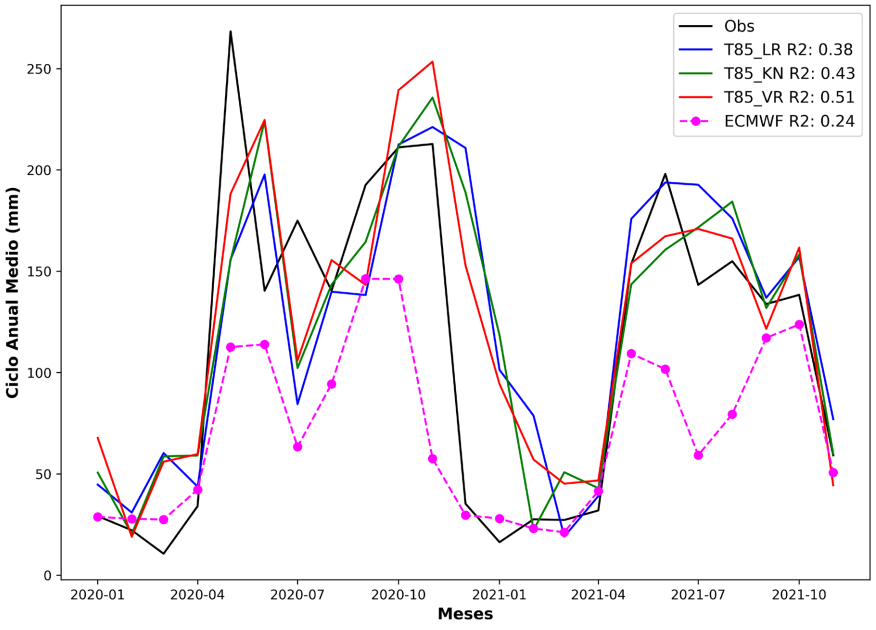



(b)

**Figure 6.** Example of forecasts resulting from 20-year sequential training to obtain monthly forecasts. (a) Caribbean SST; (b) SLP.




(a)




(b)

**Figure 7.** Example of forecasts resulting from 6-months sequential training to obtain monthly forecasts. (a) Atlantic SST; (b) Temperature at 850 hPa.

The forecasts obtained from this regression scheme further suggest that there is a general tendency to overestimate monthly rainfall during the dry season. This result suggests that the errors related to the magnitude of the changes in the predictors are not significantly corrected by the statistical component (**Figure 6** and **Figure 7**). This behavior seems to be related to representation described above, where ECMWF solutions overestimate the environment conductive for a rainfall increase. From a synoptical point of view, wetter frontal systems and a

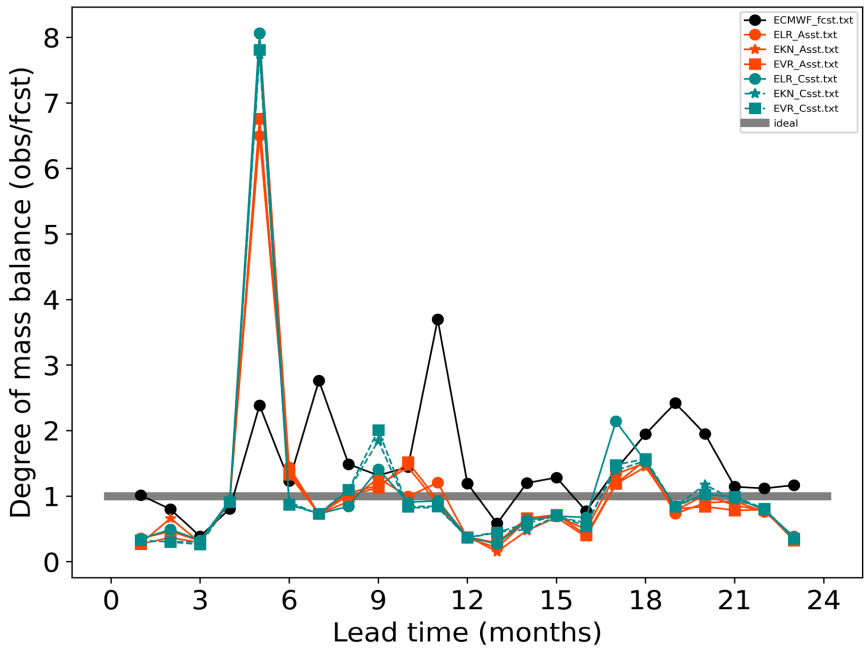
flow from the first quadrant generate precipitation episodes of variable importance, supporting this seasonal behavior.

These results expose some deficiencies that a sequential training scheme could have. In this case, the statistical models reproduce conditions of greater moisture or conductive flows to generation of different rainfall regimes and are unable to reverse this trend based on observations. This fact suggests that the use of split training schemes, either individually or combined with sequential training, could be beneficial in order to try to reverse these deficiencies.

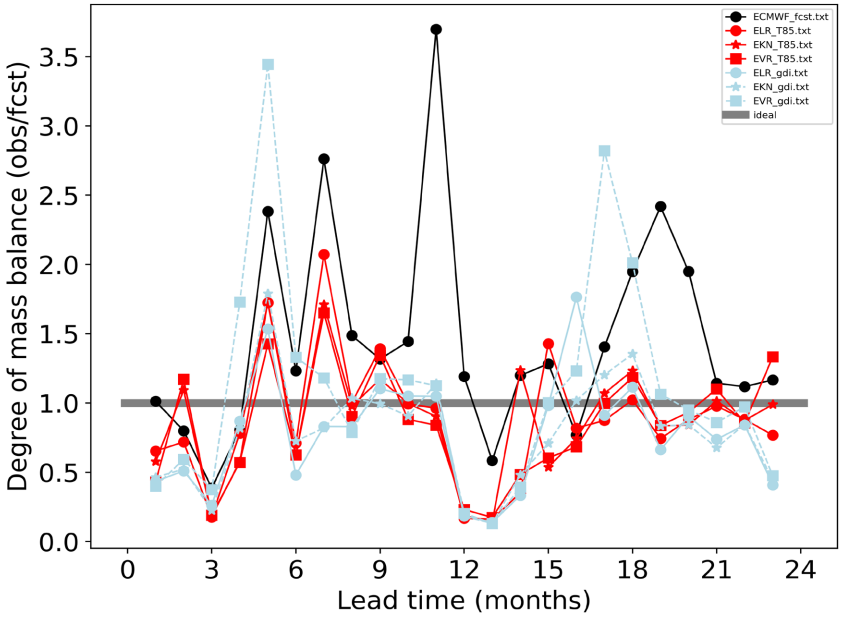
The balance schemes (**Figure 8**) show that the predictors associated with the SST had a very similar behavior, with overestimation in the winter season while exhibiting a reduction in the time scale of the wet season, shortening its duration and delaying the rainfall early peak, which leads to a significant growth in errors.

At this point, it is worth remembering that [12] suggests that warm SST biases also cover the JJA months (June-July-August). This means that the dynamical forecast can show a greater homogeneity in the tropical oceans, which from a dynamic point of view can delay the appearance of hot pools, which together with the expansion of subtropical ridge and the northward ITCZ migration, constitute a mechanism that favors convergence, moisture transport and, therefore, precipitation. This effect seems to be "absorbed", so to speak, by the models built from the SSTs, which is reflected in the results.

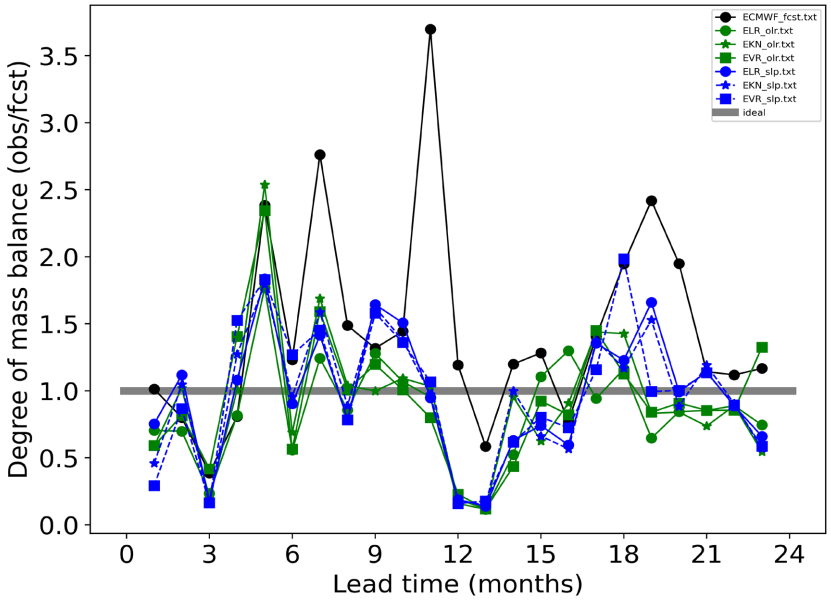
The models built with the use of the GDI as a predictor worked better when they were adjusted to a linear relationship, which makes sense since the value of the index follows this type of relationship with respect to the observational framework used (**Figure 9**). However, the nearest neighbor method showed a higher coefficient of determination than the rest (0.35), which is why it seems more convenient to use this model with the GDI.




(a)




(b)




(c)

**Figure 8.** Balance obtained between observations and design models. (a) SSTs; (b) T850 y GDI; (c) OLR y SLP.

The T850 predictor was one of those that exhibited the best results since it partially corrected the underestimation of the dynamic model during the wet season, however, the monthly rainfall overestimation in the dry season, fundamentally over the first trimester, persisted. In this case, the models built from the k-nearest neighbors and vector support methods were the most robust, with the highest determination coefficients (0.43 and 0.51 respectively), almost doubling in skill the result obtained by the ECMWF (0.23).

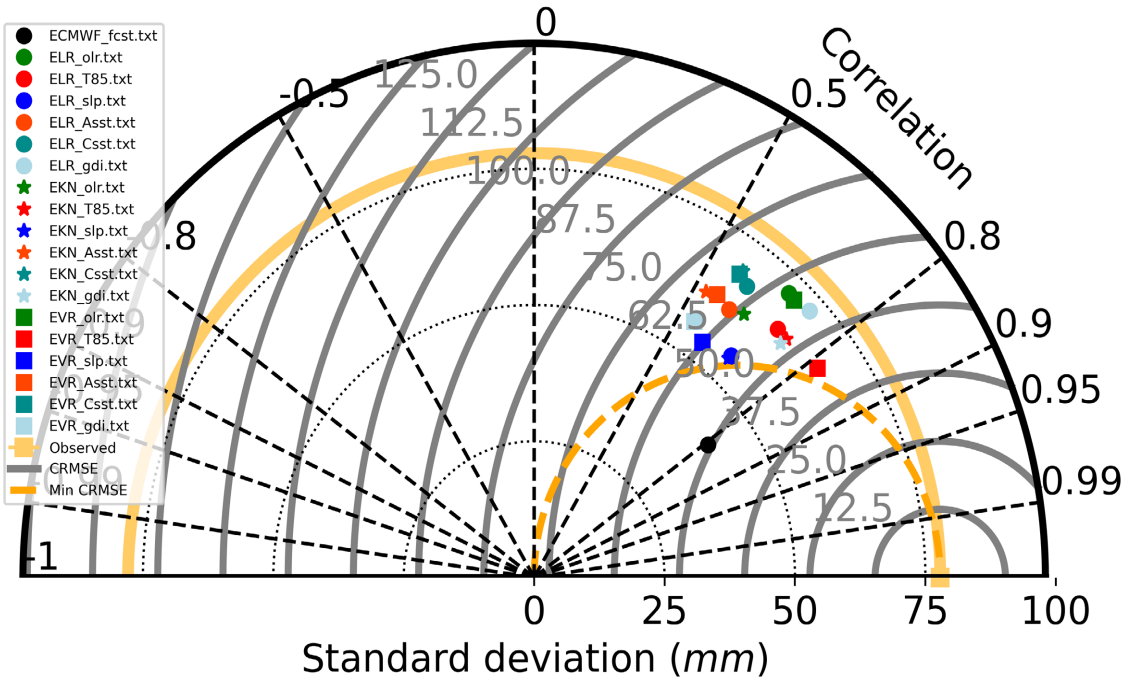


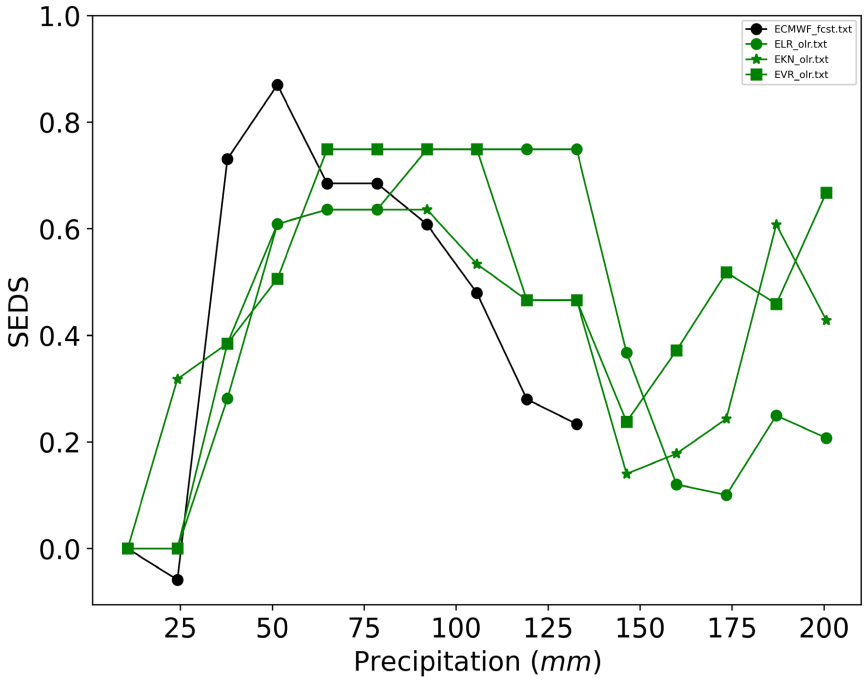

**Figure 9.** Taylor diagram including the ECMWF ensemble mean forecast and proposed models for the assessment period.

These models generate SEDS values around 0.5 or higher for thresholds greater than 50 mm, which places it together with the OLR in one of the predictors with the greatest forecast skill (**Figure 10(a)**). Precisely the latter shows similar results, but it tends to show a more pronounced delay of the early peak of precipitation more evident compared to T850, with the linear regression model being the one that exhibits the most discrete results.

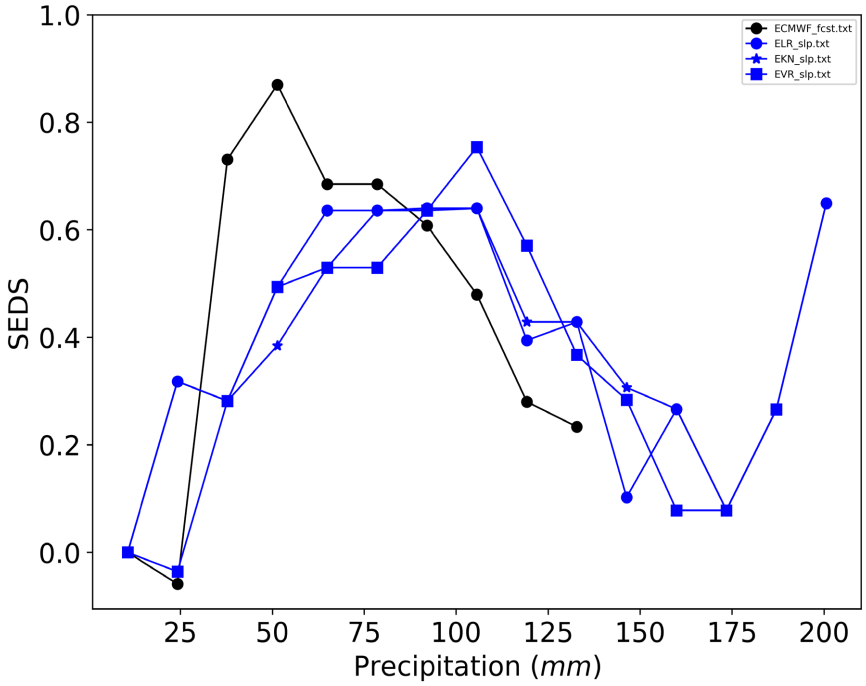
As for the SLP predictor, it represents with good approximation the monthly rainfall thresholds between 50 and just over 100 mm (**Figure 10(b)**). However, the skill indices decrease for values outside this threshold, given that their forecasts overestimate those accumulated below 50 mm and underestimate those greater than monthly values between 125 and 150 mm. In this sense, the model built from the support vectors exhibits the most realistic forecasts.

In summary, it is observed that the models built from SSTs are not suitable for monthly rainfall forecasting, taking into account these experiments designs. The SLP and GDI predictors present slightly better results, but limitations persist in relation to the forecast thresholds, where the greatest skills are obtained around the forecast mean with evident problems of overestimation/underestimation towards the lower/upper limits of forecasted series. In these cases, the models built with the k-nearest neighbor method are the most realistic.

The OLR and T850 predictors are those that represent the monthly accumulated rainfall with greater skill, although, like the rest of the predictors, they overestimate accumulations below the 50 mm threshold. Here the linear regression models presented the least satisfactory results.




(a)




(b)

**Figure 10.** Symmetric extreme dependency index calculated for the models built with the OLR (a) and SLP (b). ECMWF forecasts (black).

### 3.3. Ensemble Forecast

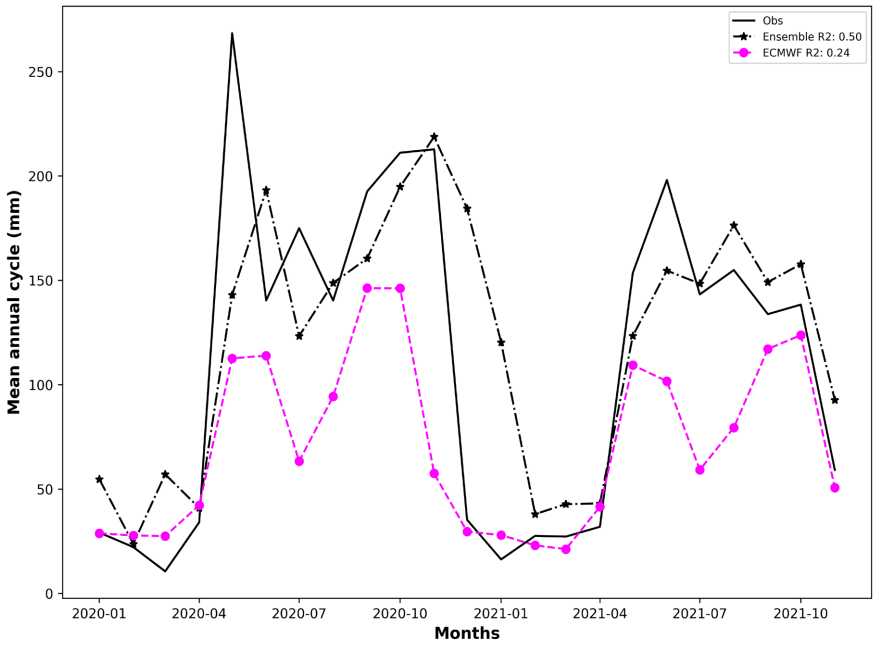
Based on described results in the previous section, it is proposed to build an ensemble with the best performing models in order to take the individual advan-

tage of each one. Those statistical-dynamic models with the lowest average bias and the highest coefficient of determination with respect to the ECMWF were selected to form the ensemble.

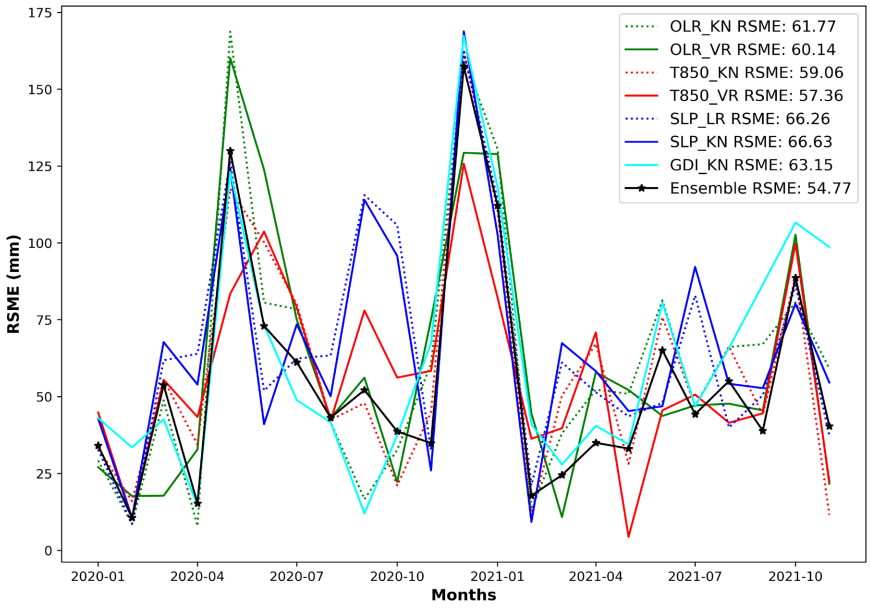
Following this philosophy, the models built from the k-nearest neighbor methodology with the predictors GDI, SLP, T850 and OLR, together with simple linear regression using the SLP predictor and the vector support methodology with T850 and OLR constitute the ensemble members.

The results suggest that the forecasts using the ensemble mean allow capturing the seasonal characteristics of the wet season, significantly correcting the underestimation present in the dynamic model and, in this case partially, the lag present in some individual models in relation to the appearance of the early rainfall peak (**Figure 11(a)**). In the transition months (spring and autumn), it also achieves some improvement with respect to the individual ensemble members, here the vector support model building with the T850 predictor, showed the lowest average errors in this period, where the atmospheric dynamic is usually complex, suggesting that T850 is an excellent predictor for this region. This is coincident with theoretical considerations because this variable represents the moisture retention capacity at lower levels, being, in fact, a good index for rainfall availability in tropical regions.

Given that the ensemble average considers the forecast values of each individual model, no significant improvements are obtained in relation to the accumulated overestimation present in the DJF trimester, this being the lowest skill point of the forecast in general. Despite these shortcomings, the ensemble forecast manages to reduce the forecast errors with respect to each individual model and the ECMWF (**Figure 11(b)**).



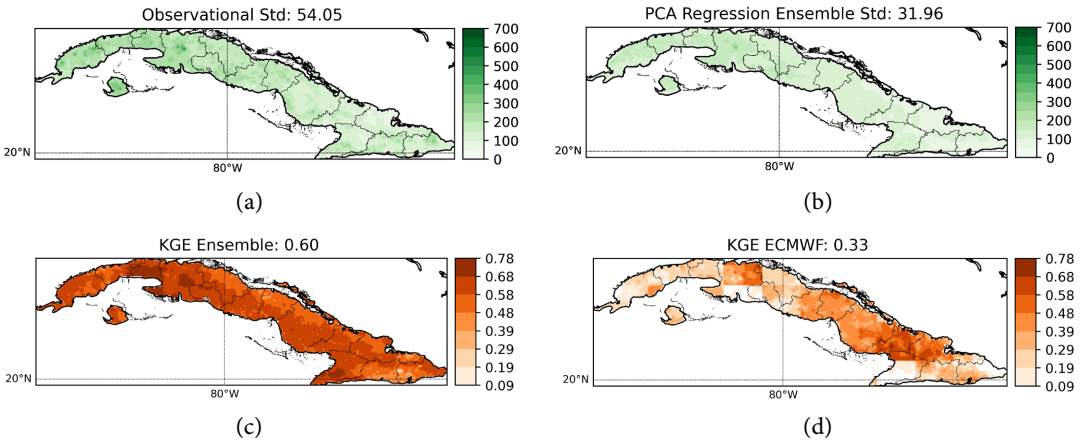

(a)




(b)

**Figure 11.** (a) Comparison of annual cycle forecast for the selected years between the mean of the statistical-dynamic ensemble and ECMWF. (b) RSME corresponding to the mean of the statistical-dynamic ensemble and each of its members.

From a spatial point of view, the ensemble building adequately reflects the rainfall spatial distribution throughout the island, even in those months where quantitative errors are usually greater, with statistical distributions that do not differ significantly from observations (**Figure 12**). This suggests that the proposed ensemble adequately captures the spatio-temporal rainfall variability but may inadequately reflect the thresholds from a quantitative point of view, as it is affected by errors inherent to the dynamic model. The application of bias correction methodologies constitutes other options that could improve the forecast quality.




(a) (b) (c) (d)

**Figure 12.** Comparison between the monthly rainfall records (a) and the forecasts resulting from the statistical-dynamic ensemble mean (b) corresponding to the months of June; and KGE value taking into account the study period between the mean of the statistical-dynamic ensemble (c) and ECMWF (d).

Compared to the ECMWF, the use of the statistical-dynamic ensemble forecast with a higher skill the rainfall seasonal behavior in the island, taking into account the period analyzed, which is reflected in average KGE values that are up to 90% higher (**Figure 12**). This improvement takes into account a better fit in relation to spatial resolution, since the proposed models, when trained in a high-resolution observational framework, can capture details of seasonal variability with a higher degree of discrimination than the European model.

## 4. Conclusions

The present research proposes to apply different PCA regression schemes through a group of previously identified predictor variables, extracted from the solutions of the ECMWF-SEAS5 ensemble mean, with the purpose of obtaining quantitative rainfall forecasts for one month.

The evaluation of the proposed models suggests that the use of sequential training schemes with periods of 20 years does not lead to realistic results in any case. In contrast, the application of this training philosophy to periods of 6 months allows for generating forecasts that reflect the rainfall seasonal trend on the island.

The study leads to the use of T850, OLR and SLP in that order being the predictors that lead to better results. The introduction of the GDI as a predictor leads to realistic forecasts, in this sense more extensive research with the index can provide a better under-standing for predictive purposes for the area taking into account its characteristics.

Of the proposed models, the k-nearest neighbor methodology generated results with higher skill indices in most predictors, except for T850 and OLR where the vector support models were more efficient. The main difficulties observed in the forecasts are related to the overestimation of the monthly rainfall in the dry season, a situation that was more marked in the DJF trimester, as well as a delay in the prediction of the early rainfall peak. Rainfall's monthly thresholds under 50 mm and above 150 mm (to a lesser degree) show lesser performance.

Building an ensemble with the individual models the quality with respect to the individual predictions of the members is enhanced, exceeding the $R^2$ value of the individual models by approximately 0.20 and the ECMWF by 0.25. However, since no element of weighting or bias correction is included, the proposed ensemble fails to significantly improve the prediction deficiencies observed in the DJF trimester.

The proposal presented in this research manages to better spatially and quantitatively represent the accumulated monthly precipitation in a general sense expressed in several metrics, as mean values of KGE and $R^2$ are 90% and just over 50% higher than those of the ECMWF.

Extending the study period, the application of this methodology to supervised learning schemes as well as evaluating other data sources for the predictors will allow refining the construction of these models in order to improve predictions.

## Acknowledgements

The authors of this research are grateful for the support provided through the project "Building Resilience to Drought in Cuba".

This research was carried out thanks to a grant from the International Development Research Center (IDRC), Ottawa, Canada.

The opinions expressed do not necessarily represent those of IDRC or its Board of Governors.

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

[1] Walker, G.T. (1910) Correlation in Seasonal Variations of Weather, II. *Memoirs of the Indian Meteorological Department*, **21**, 22-30.

[2] Walker, G.T. (1914) A Further Study of Relationships with Indian Monsoon Rainfall, II. *Memoirs of the Indian Meteorological Department*, **23**, 123-129.

[3] Walker, G.T. (1923) Correlation in Seasonal Variations of Weather, VIII. A Preliminary Study of World Weather. *Memoirs of the Indian Meteorological Department*, **24**, 75-131.

[4] Bjerknes, J. (1966) A Possible Response of the Atmospheric Hadley Circulation to Equatorial Anomalies of Ocean Temperature. *Tellus*, **18**, 820-829. https://doi.org/10.1111/j.2153-3490.1966.tb00303.x

[5] Bjerknes, J. (1969) Atmospheric Teleconnections from the Equatorial Pacific. *Monthly Weather Review*, **97**, 163-172. https://doi.org/10.1175/1520-0493(1969)097<0163:atftep>2.3.co;2

[6] Madden, R.A. and Julian, P.R. (1971) Detection of a 40-50 Day Oscillation in the Zonal Wind in the Tropical Pacific. *Journal of the Atmospheric Sciences*, **28**, 702-708. https://doi.org/10.1175/1520-0469(1971)028<0702:doadoi>2.0.co;2

[7] Madden, R.A. and Julian, P.R. (1972) Description of Global-Scale Circulation Cells in the Tropics with a 40-50 Day Period. *Journal of the Atmospheric Sciences*, **29**, 1109-1123. https://doi.org/10.1175/1520-0469(1972)029<1109:dogscc>2.0.co;2

[8] Guidance on Operational Practices for Objective Seasonal Forecasting. World Meteorological Organization (WMO-No 1246) 2020 Edition. https://library.wmo.int/records/item/57090-guidance-on-operational-practices-for-objective-seasonal-forecasting?language_id=&offset=618

[9] Laing, A. and Evans, J. (2010) Introduction to Tropical Meteorology: A Comprehensive Online and Print Textbook. Version 2a. COMET Program, University Corporation for Atmospheric Research (with Free Registration). https://www.meted.ucar.edu/tropical/textbook_2nd_edition_es

[10] Vitart, F., Buizza, R., Alonso Balmaseda, M., Balsamo, G., Bidlot, J., Bonet, A., *et al.* (2008) The New VarEPS-Monthly Forecasting System: A First Step towards Seamless Prediction. *Quarterly Journal of the Royal Meteorological Society*, **134**, 1789-1799. https://doi.org/10.1002/qj.322

[11] Molteni, F., Stockdale, T., Balmaseda, M., Balsamo, G., Buizza, R., Ferranti, L., Magnusson, L., Mogensen, K., Palmer, T. and Vitart, F. (2011) The New ECMWF

Seasonal Forecast System (System 4). No. 656, European Centre for Medium-Range Weather Forecasts.

[12] Johnson, S.J., Stockdale, T.N., Ferranti, L., Balmaseda, M.A., Molteni, F., Magnusson, L., *et al.* (2019) SEAS5: The New ECMWF Seasonal Forecast System. *Geoscientific Model Development*, **12**, 1087-1117. https://doi.org/10.5194/gmd-12-1087-2019

[13] Robertson, A.W., Vitart, F. and Camargo, S.J. (2020) Subseasonal to Seasonal Prediction of Weather to Climate with Application to Tropical Cyclones. *Journal of Geophysical Research*: *Atmospheres*, **125**, e2018JD029375. https://doi.org/10.1029/2018jd029375

[14] Dutra, E., Wetterhall, F., Di Giuseppe, F., Naumann, G., Barbosa, P., Vogt, J., *et al.* (2014) Global Meteorological Drought-Part 1: Probabilistic Monitoring. *Hydrology and Earth System Sciences*, **18**, 2657-2667. https://doi.org/10.5194/hess-18-2657-2014

[15] Carrão, H., Naumann, G., Dutra, E., Lavaysse, C. and Barbosa, P. (2018) Seasonal Drought Forecasting for Latin America Using the ECMWF S4 Forecast System. *Climate*, **6**, Article No. 48. https://doi.org/10.3390/cli6020048

[16] Barnston, A.G., Thiao, W. and Kumar, V. (1996) Long-Lead Forecasts of Seasonal Precipitation in Africa Using CCA. *Weather and Forecasting*, **11**, 506-520. https://doi.org/10.1175/1520-0434(1996)011<0506:llfosp>2.0.co;2

[17] Ashby, S.A., Taylor, M.A. and Chen, A.A. (2005) Statistical Models for Predicting Rainfall in the Caribbean. *Theoretical and Applied Climatology*, **82**, 65-80. https://doi.org/10.1007/s00704-004-0118-8

[18] Martínez, C. (2021) Seasonal Climatology, Variability, Characteristics, and Prediction of the Caribbean Rainfall Cycle. https://www.researchgate.net/publication/361266119_Seasonal_Climatology_Variability_Characteristics_and_Prediction_of_the_Caribbean_Rainfall_Cycle

[19] Martinez, C., Goddard, L., Kushnir, Y. and Ting, M. (2019) Seasonal Climatology and Dynamical Mechanisms of Rainfall in the Caribbean. *Climate Dynamics*, **53**, 825-846. https://doi.org/10.1007/s00382-019-04616-4

[20] Ávila, L., Silveira, R., Campos, A., Rogiski, N., Freitas, C., Aver, C., *et al.* (2023) Seasonal Streamflow Forecast in the Tocantins River Basin, Brazil: An Evaluation of ECMWF-SEAS5 with Multiple Conceptual Hydrological Models. *Water*, **15**, Article No. 1695. https://doi.org/10.3390/w15091695

[21] Ferreira, G.W.S., Reboita, M.S. and Drumond, A. (2022) Evaluation of ECMWF-SEAS5 Seasonal Temperature and Precipitation Predictions over South America. *Climate*, **10**, Article No. 128. https://doi.org/10.3390/cli10090128

[22] Zhao, T., Schepen, A. and Wang, Q.J. (2016) Ensemble Forecasting of Sub-Seasonal to Seasonal Streamflow by a Bayesian Joint Probability Modelling Approach. *Journal of Hydrology*, **541**, 839-849. https://doi.org/10.1016/j.jhydrol.2016.07.040

[23] Hadi, S. and Tombul, M. (2018) Long-Term Spatiotemporal Trend Analysis of Precipitation and Temperature over Turkey: Spatiotemporal Precipitation and Temperature Trend over Turkey. *Meteorological Applications*, **25**, 445-455. https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.1712

[24] Elbasheer, M.E.E.E., Corzo, G.A., Solomatine, D. and Varouchakis, E. (2022) Machine Learning and Committee Models for Improving ECMWF Subseasonal to Seasonal (S2S) Precipitation Forecast.

[25] Cárdenas, P.A., Centella, A. and Naranjo, L. (1995) Teleconnection Pacific Caribbean ENSO and QBO as Forcing Climate Variability Elements. *Proceeding 6th In-*

*terannual Meeting of Statistical Climatology*, Galway, 19-23 June 1995.

[26] Cárdenas, P.A. (1999) Pronósticos mensuales de lluvias en Cuba, un modelo con varios meses de adelanto. *Revista Cubana de Meteorología*, **6**, 47-51.

[27] Álvarez-Escudero, L., Mayor, Y.G., Borrajero-Montejo, I. and Bezanilla-Morlot, A. (2021) Assessing the Potential of a Long-Term Climate Forecast for Cuba Using the WRF Model. *Environmental Sciences Proceedings*, **8**, Article No. 44. https://doi.org/10.3390/ecas2021-10338

[28] González-Jardines, P.M., Sierra-Lorenzo, M., Ferrer-Hernández, A.L. and Bezanilla-Morlot, A. (2023) Evaluation of Candidate Predictors for Seasonal Precipitation Forecasting. *Atmospheric and Climate Sciences*, **13**, 539-564. https://doi.org/10.4236/acs.2023.134031

[29] Manzanas, R., Lucero, A., Weisheimer, A. and Gutiérrez, J.M. (2017) Can Bias Correction and Statistical Downscaling Methods Improve the Skill of Seasonal Precipitation Forecasts? *Climate Dynamics*, **50**, 1161-1176. https://doi.org/10.1007/s00382-017-3668-z

[30] Gálvez, J.M. and Davison, M. (2016) The Gálvez-Davison Index for Tropical Convection. https://www.wpc.ncep.noaa.gov/international/gdi/GDI_Manuscript_V20161021.pdf

[31] Muñoz, Á.G., Yang, X., Vecchi, G.A., Robertson, A.W. and Cooke, W.F. (2017) A Weather-Type-Based Cross-Time-Scale Diagnostic Framework for Coupled Circulation Models. *Journal of Climate*, **30**, 8951-8972. https://doi.org/10.1175/jcli-d-17-0115.1

[32] Alfaro, E.J., Chourio, X., Muñoz, Á.G. and Mason, S.J. (2017) Improved Seasonal Prediction Skill of Rainfall for the Primera Season in Central America. *International Journal of Climatology*, **38**, e255-e268. https://doi.org/10.1002/joc.5366

[33] Solano Ojeda, O., Vázquez Montenegro, R. and Martín Padrón, M.E. (2006) Estudio de la extensión superficial anual de la sequía agrícola en Cuba durante el período 1951-1990. *Revista Cubana de Meteorología*, **13**, 41-52. http://rcm.insmet.cu/index.php/rcm/article/view/404

[34] Clarke, B.R. (2008) Linear Models: The Theory and Application of Analysis of Variance. John Wiley & Sons, Inc., 272 p. https://www.wiley.com/en-au/Linear+Models%3A+The+Theory+and+Ap-plication+of+Analysis+of+Variance-p-9780470025666

[35] Scikit-Learn User Guide, Release 0.18.2 (2017). https://scikit-learn.org/0.18/_downloads/scikit-learn-docs.pdf

[36] Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., *et al.* (2013) API Design for Machine Learning Software: Experiences from the Scikit-Learn Project. *European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases*, Prague, September 2013, hal-00856511 https://inria.hal.science/hal-00856511/document

[37] Ferro, C.A.T. and Stephenson, D.B. (2011) Extremal Dependence Indices: Improved Verification Measures for Deterministic Forecasts of Rare Binary Events. *Weather and Forecasting*, **26**, 699-713. https://doi.org/10.1175/waf-d-10-05030.1