Research Article

Spatiotemporal Forecasting in Climate Data Using EOFs and Machine Learning Models: A Case Study in Chile

Mauricio Herrera[1], Francisca Kleisinger[1], Andrés Wilson[1]

1. Faculty of Engineering, Universidad del Desarrollo, Chile

Effective resource management and environmental planning in regions with high climatic variability, such as Chile, demand advanced predictive tools. The success in these areas heavily relies on accurately interpreting and forecasting climatic patterns. This study addresses these challenges by employing an innovative and computationally efficient hybrid methodology that integrates machine learning (ML) methods for time series forecasting with established statistical techniques. The spatiotemporal data undergo decomposition using time-dependent Empirical Orthogonal Functions (EOFs), denoted as $\phi_k(t)$, and their corresponding spatial coefficients, $\alpha_k(s)$, to reduce dimensionality. Wavelet analysis provides high-resolution time and frequency information from the $\phi_k(t)$ functions, while neural networks forecast these functions within a medium-range horizon h. By utilizing various ML models, particularly a Wavelet–ANN hybrid model, we forecast $\phi_k(t+h)$ up to a time horizon h, and subsequently reconstruct the spatiotemporal data using these extended EOFs.This methodology is applied to a grid of climate data comprising 6355 points covering the entire territory of Chile. It transitions from a high-dimensional multivariate spatiotemporal data forecasting problem (involving 6355 time series) to a low-dimensional univariate time series forecasting problem (requiring only a few dozen forecasts). Additionally, cluster analysis with Dynamic Time Warping for defining similarities between rainfall time series, along with spatial coherence and predictability assessments, has been instrumental in identifying geographic areas where model performance is enhanced. This approach also elucidates the reasons behind poor forecast performance in regions or clusters with low spatial coherence and predictability. By utilizing cluster medoids, the forecasting process becomes more practical and efficient. This compound approach significantly reduces computational complexity while generating forecasts of reasonable accuracy and utility.

Significance Statement. The approach outlined in this study facilitates the transition from a high–dimensional multivariate spatiotemporal data forecasting problem to a low-dimensional univariate time series forecasting problem. This transition substantially reduces computational complexity while yielding

reasonably accurate forecasts and enhances our ability to interpret and predict climatic patterns across the entire territory and over medium-term temporal horizons, despite its high climatic variability.

Corresponding author: Mauricio Herrera, mherrera@udd.cl

1. Introduction

Climate change's growing complexity and urgency demand refined yet practical prediction tools, especially in vulnerable regions like Chile with high climatic variability. Effective resource management and environmental planning hinge on our ability to decipher and anticipate climatic patterns, directly impacting agricultural planning and water management. Additionally, understanding precipitation, temperature, and other climatic variables shapes crucial policies concerning climate change and environmental protection. Data-driven analysis guides decision-making towards effective mitigation and adaptation strategies[1][2].

However, the intricate spatial, temporal, and spatiotemporal correlations in environmental data pose a significant challenge in capturing these dependencies[3]. Understanding and modeling these patterns are crucial for advancing climate research and prediction. Fortunately, a convergence of interests and expertise is emerging. Climate researchers, particularly those in numerical weather prediction, atmospheric physics, extreme events, and climate change, are increasingly turning to Machine Learning (ML) to enhance modeling and prediction[4]. Similarly, ML researchers recognize the relevance of their work in addressing climate challenges, especially in numerical weather prediction[5][6][7]. This collaboration has the potential to unlock ML's capabilities for modeling complex dynamical systems in climate science.

This study employs a hybrid approach for spatiotemporal climate data analysis and forecasting. We harness empirical orthogonal functions (EOFs) for dimensionality reduction, wavelet analysis for high-resolution time-frequency information, and deep neural networks (DNNs) for forecasting. This combined approach capitalizes on the strengths of each technique to tackle the complexities inherent in climate data analysis.

Pioneered in meteorology by[8], EOF analysis has become a cornerstone for understanding spatiotemporal climate variability. As explored in seminal works like[9], EOFs reveal orthogonal patterns of variability, each explaining distinct portions of data variance. Notably, the first EOF captures the most significant variance, followed by subsequent ones with diminishing contributions. This inherent efficiency – retaining key information while minimizing complexity – makes EOFs ideal for empirical climate modeling.

The use of Empirical Orthogonal Functions (EOFs) has limitations, as they often lack clear physical significance[10]. Despite their utility in capturing dominant variance patterns, EOFs may not inherently

represent distinct physical processes, making it challenging to distinguish empirical modes from genuine physical phenomena[11]. Despite these challenges, EOFs remain valuable for making reasonable predictions with a limited number of data modes. By distilling spatiotemporal variability into a manageable set of orthogonal patterns, EOFs enable efficient empirical modeling and prediction, demonstrating their practical importance in forecasting massive spatiotemporal data. In our study, we utilize EOF analysis alongside ML and wavelets mainly for forecasting, without aiming to distinguish data–driven modes from physically meaningful structures.

This study utilizes a grid comprising 6355 points at a resolution of 0.25×0.25 degrees, covering the entirety of Chile. Each point is associated with a time series spanning from 1980 to 2022, encompassing climatic variables such as daily accumulated precipitation, maximum, mean, and minimum daily temperature, evapotranspiration, etc.

By employing Singular Value Decomposition (SVD), these datasets undergo factorization into Empirical Orthogonal Functions (EOFs) that encapsulate temporal information $\phi_k(t)$ (note that we use empirical temporal orthogonal functions, since they are obtained from an empirical temporal covariance matrix) and the corresponding spatial coefficients $\alpha_k(s)$, which capture spatial information[12].

A Wavelet–ANN Hybrid model for forecasting, constructed upon wavelet transform using the *Maximal Overlap Discrete Wavelet Transform* (MODWT) algorithm developed by[13], facilitates the forecasting of EOFs $\phi_k(t+h)$ over a horizon h. The spatiotemporal data is then reconstructed utilizing this extended temporal component over a horizon h.

This methodology facilitates the transition from a high-dimensional multivariate spatiotemporal data forecasting problem (in this case, entailing forecasting using 6355 time series corresponding to grid points) to a low-dimensional univariate time series forecasting problem (in this case, up to a couple of dozens of forecasts), significantly reducing computational complexity while yielding forecasts of reasonable utility.

To account for the extensive climatic variability of Chile, we use cluster analysis based on time series. We group time series of rainfall with manifest similarities measured by distances based on *Dynamic Time Warping* (DTW). The structure of clusters or geographic segmentation of similar rainfall patterns is very stable over time.

To make forecast with the proposed methodology, we use the medoids of each cluster, making forecasting in each geographic zone more practical and effective. We conduct these forecasting tests to demonstrate that the proposed methodology has accurately captured the patterns of precipitation behavior over time and space. Additionally spatial coherence and predictability assessments for each clusters, has been instrumental in identifying geographic areas where model performance is enhanced. This approach also elucidates the reasons behind poor forecast performance in regions or clusters with low spatial coherence and predictability.

2. Materials and methods

a. Data

The data were obtained from ERA5 in[14], which is part of the Copernicus Climate Change Service (C3S) provided by the European Union and produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). ERA5 offers reanalyzed climatic and meteorological data, covering the period from 1950 to the present, providing detailed information on a wide range of atmospheric, terrestrial, and oceanic variables. It is known for its high spatial resolution and temporal resolution, making it widely used in climate research, environmental studies, and meteorological modeling applications.

The data used in this study form a grid of 6355 points located between the geographical coordinates - latitude -17.5° to -56.0° and longitude -76.0° to -66.0° with resolution of 0.25×0.25 degrees. Each point is associated with historical data containing time series of climatic variables such as maximum, mean, and minimum temperature, precipitation, evapotranspiration, etc. Additionally, each point is characterized by its geographical coordinates – longitude and latitude.

Let $X(s,t) = \{x(s_i, t_j)\}$ be a data structure for a spatiotemporal variable x (e.g. precipitation, temperature, etc.) with $i = 1...n$ and $s_i = (\text{long}_i, \text{lat}_i, \text{e}_i$. Where $i = 1..., n$, long_irepresents longitude, lat_i represents latitude, and e_i represents elevation of a location i (see Table 1). So, $X(s,t)$ consists of n locations, each having an associated time series with p records. In this study, we consider data corresponding to 41 longitude values and 155 latitude values, creating a grid of $n = 6355$ spatial points. For each of them, there are $p = 26665$ time values, corresponding to daily records between 1980 and 2022.

	t_1	t_2	$\cdots$	t_p
s_1	x_{s_1,t_1}	x_{s_1,t_2}	$\cdots$	x_{s_1,t_p}
s_2	x_{s_2,t_1}	x_{s_2,t_2}	$\cdots$	x_{s_2,t_p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
s_n	x_{s_n,t_1}	x_{s_n,t_2}	$\cdots$	x_{s_n,t_p}

Table 1. Tidy structure of spatiotemporal data for analysis.

b. Decomposition of spatiotemporal data using EOFs.

Let $\mathcal{I}_n$ be a column matrix with n ones. With this matrix, the time mean is expressed as $\bar{x} = \frac{1}{n} X^T \mathcal{I}_n$.

The empirical temporal correlation matrix with dimensions $p \times p$ is written as

$$C_{p\times p} = \frac{1}{n} X^T X - \bar{x}\bar{x}^T = \frac{1}{n} X^T H X$$

Where $H = \mathbb{I} - \frac{1}{n}\mathcal{I}_n \mathcal{I}_n^T$ is the so called *centering matrix*.

We transform the data table 1 by subtracting the time mean from each value using $Y = X - \mathcal{I}_n \bar{x}^T$. So, $\bar{y} = 0$ and:

$$C_{p\times p} = \frac{1}{n} Y^T Y$$

Introducing $Z = \frac{Y}{\sqrt{n}}$ (spatially centered and normalized data) $\Rightarrow C_{p\times p} = Z^T Z$.

Instead of directly calculating the eigenvalues and eigenvectors of the matrix $C_{p\times p}$, the matrix Z is generally (more efficiently) factorized using Singular Value Decomposition (SVD). $Z_{n\times p} = U_{n\times n} D_{n\times p} V^T_{p\times p}$. Therefore, $Z^T Z = V D^T D V^T$. Comparing with the spectral decomposition of the matrix $C_{p\times p} = \Phi \Lambda \Phi^T$, where Φ is the matrix of eigenvectors and Λ is a matrix with the eigenvalues of $C_{p\times p}$ on the diagonal, we have $\Phi = V$, $D^T D = \Lambda$, and the Principal Components (PCs) are $\alpha = UD$ [15].

The EOFs can be defined as the eigenvectors of covariance matrix $C_{p\times p}$.

$$C_{p\times p}\phi_k = \lambda_k \phi_k$$

As $C_{p\times p}$ is a nonnegative definite square matrix, the eigenvalues λ_k are all nonnegative and the eigenvectors ϕ_k form a complete orthonormal basis. So, the centred and normalized data $Z(s,t) = \{z(s_i, t_j)\}$ can be represented using a discrete temporal orthonormal basis $\{\phi_k(t_j)\}_{k=1}^{k=K}$ as:

$$z_{s_i,t_j} = \sum_{k=1}^{K} \alpha_k(s_i)\phi_k(t_j), \quad i = 1,\ldots,n; \quad j = 1,\ldots,p \tag{1}$$

Where $K = \min(n,p)$ and $\alpha_k(s_i)$ is the coefficient corresponding to the k –th basis function $\phi_k(t_j)$ at spatial location s_i. It is noteworthy that the scalar coefficient $\alpha_k(s_i)$ depends solely on the location and not on time, whereas the temporal basis function $\phi_k(t_j)$ is independent of space. The rationale behind this decomposition is theoretically grounded in the Karhunen-Loève expansion[9].

The coefficients α_k in this expansion can be calculated using

$$\alpha_k = Z \cdot \phi_k$$

Alternatively, using matrices

$$\alpha = ZV = UD$$

So, the expansion coefficients $\alpha_k(s_i)$ are the spatial Principal Components (PCs).

It follows from the fact that the ϕ_k are orthonormal eigenvectors of $C_{p\times p}$ that the PC are mutually uncorrelated and:

$$E(\alpha_k(s_i)) = 0$$

$$Var[\alpha_1(s_i)] \geq Var[\alpha_2(s_i)] \geq \cdots \geq Var[\alpha_K(s_i)] \geq 0$$

$$Cov[\alpha_{k_1}(s_i), \alpha_{k_2}(s_i)] = 0 \quad \text{for all } k_1 \neq k_2,$$

So, considering the SVD, the functions ϕ_k are given by V_k, where V_k represents the k-th column of the matrix V in the SVD of Z. The normalized spatial coefficients are $\alpha_k = Z \cdot V_k = U_k D$.

The—potentially truncated—EOFs decomposition ($\bar{K} < K$) returns for each spatial location s_i, corresponding to the original observations, $\bar{K}$ random coefficients α_k. These coefficients can be spatially modelled and mapped on a regular grid solving an interpolation/regression task[16].

c. Forecasting EOFs

In this study, we forecast the EOFs $\phi_k(t)$, in principle, to an arbitrary horizon h using various forecasting techniques. Specifically, the Wavelet-ANN Hybrid Model, as introduced by[13], consistently yields superior results in forecasting. Leveraging this model, we generate forecasts to predict $\phi_k(t+h)$ with a horizon h. Subsequently, utilizing this function and the spatial coefficients $\alpha_k(s)$, we reconstruct the complete spatiotemporal data. The underlying hypothesis suggests that these spatial coefficients should undergo minimal changes since significant alterations in spatial information are not anticipated within the forecast interval h (see Fig. 4 for an example).

This approximation must consider the uncertainty associated with the temporal forecast of the EOFs. Additionally, when selecting a number $\overline{K}$ of EOFs, we must also incorporate errors associated with the dimensionality reduction.

$$z_{s_i,t_j} \approx \sum_{k=1}^{\overline{K}} \alpha_k(s_i)\phi_k(t_j), \quad i = 1, \ldots, n; \quad j = 1, \ldots, p+h$$

To reconstruct the original approximated spatiotemporal data, but extended to horizon h, $\hat{X}(s, t+h)$, we have:

$$\hat{X}(s, t+h) = \sqrt{n} U_{n\times\overline{K}} D_{\overline{K}\times\overline{K}} V^T_{\overline{K}\times p+h} + \mathcal{I}_n \overline{x}^T \tag{2}$$

Where $\tilde{D}_{\overline{K}\times\overline{K}}$ is obtained from the matrix D in the SVD decomposition by using only $\overline{K} \leq K$ columns and rows, and $\tilde{V}$ is the matrix V, but taking $\overline{K}$ rows extended with h new elements from the forecast.

Note that here we use the same $\bar{x}$ under the assumption that incorporating more temporal records does not drastically alter the mean value.

d. Some Considerations for the Application of the Proposed Predictive Model

1. Potential errors in predicting the EOFs $\phi_k(t+h)$ can propagate and impact the reconstruction of spatiotemporal data. To minimize these propagated errors, it is crucial to achieve accurate predictions within an appropriate horizon h by using the most suitable predictive model.
2. If the relationship between spatial components $\alpha_k(s)$ and temporal components $\phi_k(t)$ changes significantly over the chosen prediction horizon, the assumption of using the same spatial coefficients $\alpha_k(s)$ for both $\phi_k(t)$ and the extended $\phi_k(t+h)$ may be invalid. To capture the spatiotemporal correlations and their stability, we segment the data for the entire territory using clusters. Cluster analysis, along with its assessment of spatial coherence (and thus predictability), aims to identify geographic areas where spatial coherence enhances model performance. It also explains poor prediction results in clusters with low spatial coherence and predictability. For more efficient predictions, we use the medoids of the clusters to apply the model. In clusters with low predictability, results can still be managed by reducing the prediction horizon.
3. If the original data contains nonlinearities, EOFs decomposition may not be suitable. It is advised to verify, prior to prediction, the alignment between the real data and the data reconstructed from previously decomposed EOFs. If the alignment is good, the model can be applied to achieve accurate forecasts.
4. The proposed model's predictions are valid for short to medium horizons. Using historical data with extensive records may include complex nonlinear patterns and significant variations in spatiotemporal conditions that will affect the forecast with this method. For studying historical variations and long-term changes, it is recommended to use models other than the one proposed here (for example, Non-Homogeneous Hidden Markov Models).

e. Analyzing Precipitation Pattern Similarities Using Dynamic Time Warping on Time Series

In this section, we compare precipitation time series from 460 strategically selected geographic locations from the data grid and across Chile in search of similarity patterns. These locations are positioned around actual meteorological stations nationwide[17].

The objective of time series comparison methods is to produce a distance metric between two input time series. The similarity or dissimilarity of two time series is typically calculated by converting the data into vectors and calculating the Euclidean distance between those points in vector space. Traditional time series Euclidean Matching is extremely restrictive. However, Dynamic Time Warping (DTW)[18] allows the two curves to match

up evenly even though the X-axes (i.e., time) are not necessarily in sync. The rationale behind DTW is to stretch or compress two time series locally in order to make one resemble the other as much as possible. The distance between the two is then computed, after stretching, by summing the distances of individual aligned elements.

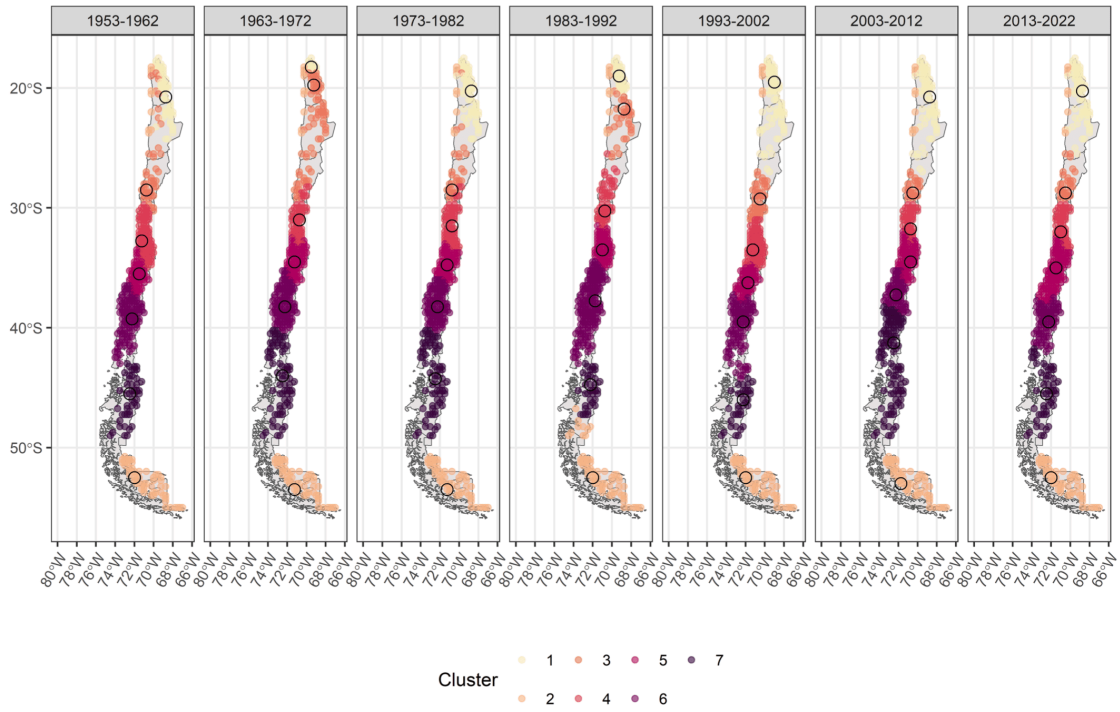

Figure 1. Solution of Seven Clusters Using Hierarchical Method and DTW – Based Distance. Records are segmented into ten-year intervals to illustrate stability versus slight variations in cluster structure.

We are employing the Dynamic Time Warping (DTW) method for a cluster analysis with two primary objectives:

1. To demonstrate that the relationship between precipitation patterns and geographical location is generally stable (stability of spatiotemporal precipitation patterns).
2. For cluster-based forecasting. By focusing on clusters, we can improve forecast accuracy for the region, for example, by using the representative (medoid) of each cluster for model predictions.

Fig. 1 presents a seven-cluster structure obtained using the hierarchical algorithm with Ward method, and with DTW-based distance. To demonstrate the stability of spatiotemporal patterns, precipitation data from 460 locations across the territory, with time series records from 1953 to 2022, are segmented into 10 – year intervals, and clusters were constructed for each interval. Only records from the months of May, June, July, and August (MJJA), which correspond to the rainy season in Chile, were considered. This segmentation aims to

capture the stability of rainfall patterns and observe any minor changes in the structure of these precipitation patterns.

Selecting a structure with seven clusters ensures a segmentation of the data that provides enough records in each cluster to proceed with EOF analysis and further training of the ML models for time series. This structure effectively captures the specific patterns present in the precipitation data, enabling a detailed and meaningful analysis.

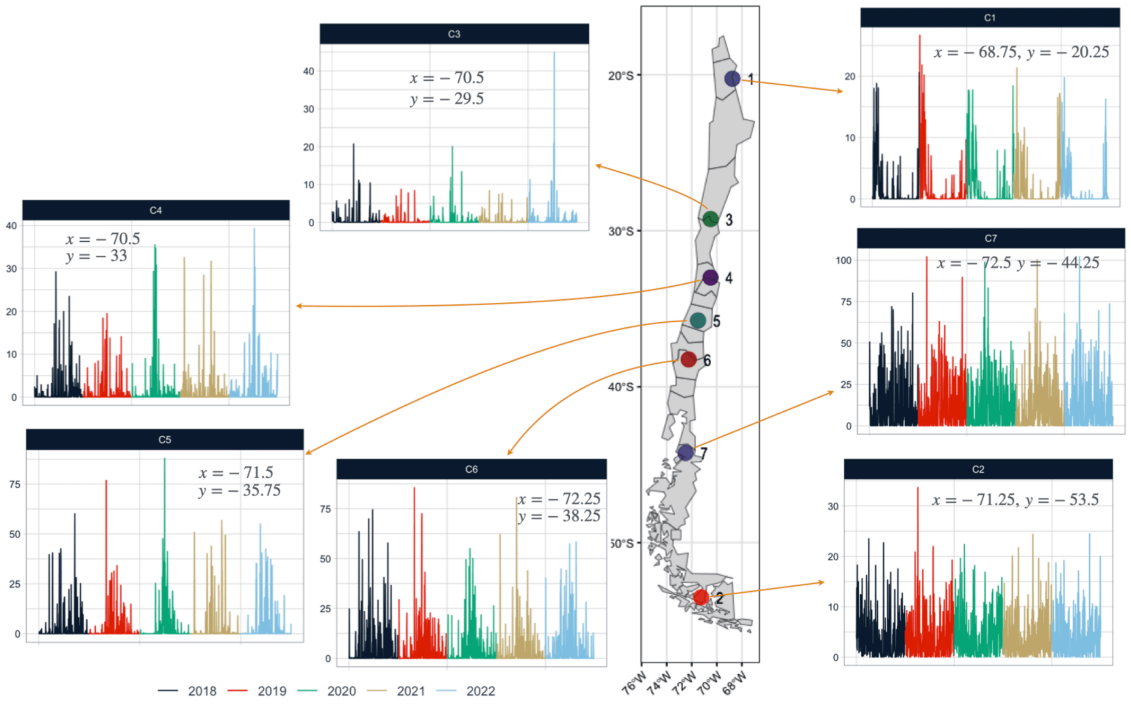

Figure 2. Cluster–Specific Precipitation Patterns from 2018 to 2022

Fig. 2 depicts the characteristic precipitation patterns of each cluster. The time series shown in the figure represent the precipitation records from 2018 to 2022 for the medoids, or typical representatives, of each cluster.

f. Spatial coherence and potential predictability.

The spatial coherence provides a measure of potential predictability at the location scale[19]. Two scores are frequently used to provide empirical estimates of the spatial coherence of seasonal anomalies between locations: $var(SAI)$ - the interannual variance of the standardized anomaly index[20], and DOF - the number of spatial degrees of freedom[21].

doi.org/10.32388/V71U3T

The SAI is computed by standardizing the interannual time series at each location (subtracting the mean and dividing by the STD) and then averaging the standardized anomalies spatially across the locations to form an index; it thus gives each location equal weight in the index. The amplitude of the SAI for a particular year depends on the size of the correlations between locations, and thus its variance gives a measure of spatial coherence of the field.

$$var(SAI_i) = var\left[\frac{1}{n}\sum_{j=1}^{n}\frac{x_{ij} - \overline{x}_j}{\sigma_j}\right]$$

where x_j is the long–term time mean over $i = 1...l$ years and σ_j is the interannual standard deviation for location j. The $var(SAI)$ is a maximum when all locations are perfectly correlated, $var(SAI) = 1$, and a minimum when the locations are uncorrelated, resulting in a $var(SAI) = \frac{1}{n}$.

The DOF gives an empirical estimate of the spatial coherence in terms of empirical (spatial) orthogonal functions, with higher values denoting lower spatial coherence:

$$DOF = \frac{n^2}{\sum_{j=1}^{n}\lambda_j^2}$$

where λ_j are the eigenvalues of the correlation matrix formed from the location seasonal–mean time series and n is the number of locations.

		DOF			var(SAI)		
	N°	**RAm**	**RI**	**RF**	**RAm**	**RI**	**RF**
Clusters							
1	64	2,70	5,94	2,70	0,52	0,29	0,54
2	64	2,58	5,57	2,47	0,51	0,26	0,53
3	48	1,39	2,58	1,81	0,84	0,59	0,73
4	75	1,20	1,65	1,46	0,91	0,77	0,82
5	53	1,16	1,51	1,21	0,93	0,81	0,91
6	80	1,25	1,81	1,34	0,89	0,73	0,86
7	76	2,05	2,97	1,92	0,64	0,49	0,69
Altitude class							
0-500 m	293	3,46	5,78	3,68	0,27	0,21	0,28
500-1500 m	87	2,56	3,66	2,90	0,50	0,39	0,49
$>$1500 m	80	3,41	6,91	3,53	0,42	0,26	0,37
All locations	460	4,11	6,68	4,56	0,24	0,20	0,22

Table 2. Table of Clusters for Degrees of Freedom (DOF) and Variance of the Standardized Anomaly Index (var(SAI)) for Accumulated Rainfall, Rainfall Intensity, and Rainfall Frequency during the MJJA period for the years 1980–2022

Table 2 displays the *DOF* and $var(SAI))$ for Accumulated Rainfall (**RAm**), Rainfall Intensity (**RI**: calculated as the total millimeters of rain divided by the number of rainy days. Where daily precipitation exceeds 1 mm), and Rainfall Frequency (**RF**: calculated as the number of rainy days divided by the total number of days in the period) from 1980 to 2022 (considering only MJJA) across 460 locations. Additionally, it considers the cluster structure based on DTW.

Fig. 3 shows the *DOF* and $var(SAI)$ metrics for RAm. The cases "All" (all locations), "$0 - 500$m" (locations with elevations between $0 - 500$m), and "> 500m" (locations with elevations above 1500 m) have the highest DOF and the lowest var(SAI) values. This indicates that the spatial coherence of these locations is lower than in other clusters. For "All" and "0-500m", this can be attributed to the large number of locations considered in the

calculation (460 and 293, respectively), resulting in a diversity of microclimates. For "$>$ 500m", the lower coherence is explained by the effect of orographic rainfall at these elevations.

On the other hand, clusters 3, 4, 5, and 6 exhibit values indicating better spatial coherence. The DOF for these clusters is close to the minimum (DOF = 1) and $var(SAI)$ is close to the maximum ($var(SAI) = 1$). These clusters correspond to the central and central–northern regions of Chile. These values indicate that locations within these clusters have very similar behavior patterns. Given the high spatial coherence in these clusters, predictive models are likely to capture their precipitation patterns accurately. However, this is not the case for clusters 1, 2, and possibly 7, which have DOF and $var(SAI)$ values indicating low spatial coherence. It is worth noting that the clustering structure found generates groups with significantly lower DOF and higher $var(SAI)$ compared to segmentations based on elevation ranges.

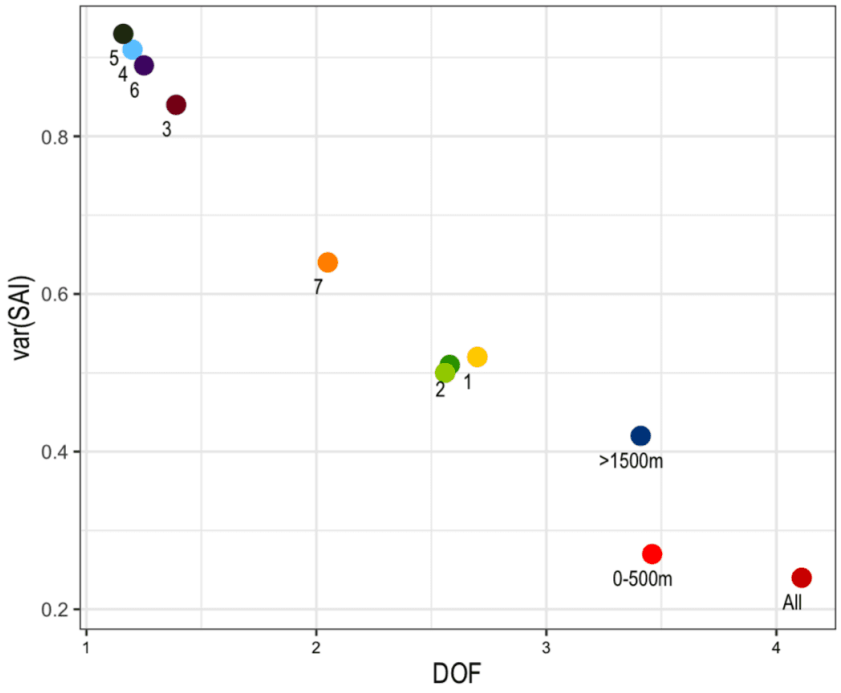

Figure 3. DOF and $var(SAI)$ metrics for Rainfall amount (RAm) by clusters (1 - 7). Groups "All", "0 – 500m", "500 – 1500m" and "$>$ 500m" are included.

Considering RI results in a general decline in both indicators' values across all clusters, with more significant effects in some. Clusters 1 and 2, which already had the lowest coherence for accumulated precipitation, show similar trends in rainfall intensity, aligning more with"All", "0 – 500m", and "$>$ 500m". This indicates greater variability in rainfall intensity within these clusters compared to average precipitation, with worsening effects

especially pronounced in clusters 1 and 2. For clusters with higher coherence, cluster 3 deteriorates and diverges from previously better–performing clusters (4, 5, and 6). This is due to its northern location, adjacent to cluster 1, both situated in the arid regions of North Grande and North Chico in Chile, which experience scarce and infrequent rainfall.

For RF, clusters 1 and 2 continue to exhibit the worst coherence, while clusters 4, 5, and 6 maintain the best.

In Summary:

- Clustering results in groups with lower *DOF* and higher $var(SAI)$ compared to separation by altitude ranges, consistent with capturing micro-climates unique to the clusters.
- Clusters 4, 5, and 6 show the best spatial coherence.
- The low coherence in rainfall intensity for cluster 1 is due to its location in northern Chile, an area with very scarce precipitation.
- High coherence in clusters 4, 5, and 6 suggests that predictive models can accurately capture their patterns and perform well.
- Low to medium spatial coherence in clusters 1, 2, 3, and 7 implies challenges in using predictive models effectively.

3. Results

a. Forecasting Precipitation Across the Chile Territory.

To test the forecast using the described method, we will consider precipitation records from a five-year period between 2018 and 2022 for all data grid points covering the country. The first four years (2018–2021) are used as training data to fit the predictive model, and the forecast is made for the entire year 2022, representing a horizon of $h = 365$ days.

Both training and forecasting are performed independently for all grid points within each cluster's defined geographic area. Recall that the clusters are constructed from a sample of 460 points (out of a total of 6355 grid points) near the precipitation measurement stations. The cluster structure segments this data sample by capturing spatiotemporal correlations. Thus, similar precipitation patterns in the records (similarity defined using DTW distances) are associated with specific geographic areas. Moreover, this cluster structure is stable enough for analysis. Here, "stable enough" means a minimal variability of the cluster structure during the period chosen for model fitting and the forecast horizon considered. These stable spatiotemporal correlations allow the use of extended EOFs for forecasting to reconstruct the data. We thus consider the forecast in seven

geographic regions, including all grid points falling within each cluster-defined region, to fit the predictive model.

Depending on the cluster, the number of data grid points (time series) considered varies, as each cluster has a different number of locations within its delineated geographic zone. Additionally, the number of EOFs to be considered also depends on the chosen cluster. The criterion is that this number of EOFs should ensure more than 80% of the explained variance. Clusters with better DOF and Var(SAI) indicators require fewer EOFs for the analysis compared to clusters with poorer indices. Thus, clusters labeled 3 to 6, which present better metrics, require between 3 to 5 EOFs to explain more than 80% of the variance, while clusters 1, 2, and 7 require more than 5 EOFs to achieve this.

We decomposed the data using Singular Value Decomposition (SVD), obtaining Empirical Orthogonal Functions (EOFs) and Principal Components (PCs). In practice, this is akin to decomposing the data into a series using EOFs as basis functions, where the coefficients represent the PCs. We selected $\overline{K} = 5$ to $\overline{K} = 7$ PCs for each cluster-defined geographic region, accounting for more than 80% of the explained variance in each case, and used their corresponding EOFs to forecast the spatiotemporal data for each region.

The coefficients $\alpha_k(s)$ (PCs), which solely depend on spatial coordinates, demonstrate minimal variability when extending the forecasting horizon to a year (i.e.,to 2022), as supported by empirical evidence. Upon meticulous examination of EOFs and their associated spatial coefficients across various temporal intervals in spatio–temporal data decomposition, it becomes evident that the first spatial coefficients (i.e., the first PCs) exhibit negligible variation across these different studied time intervals. Fig. 4 illustrates this by comparing the first 10 spatial coefficients calculated for the dataset between 2018 and 2021 with those calculated when incorporating the year 2022.

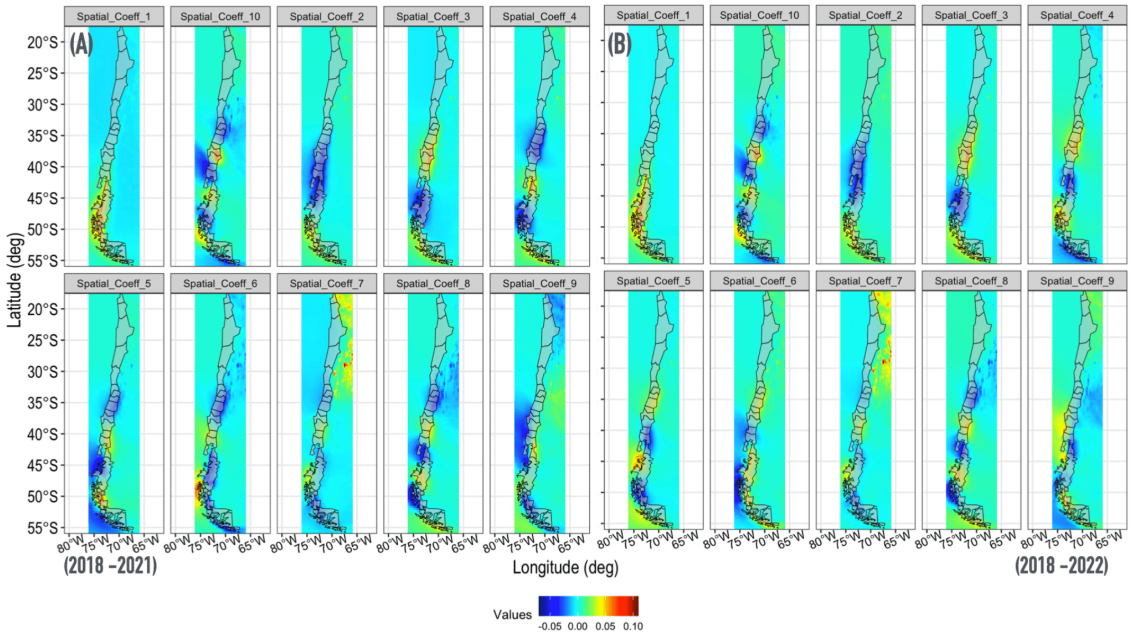

Figure 4. Comparison of the top 10 spatial coefficients from EOF decomposition of precipitation data for 2018–2021 (panel A) and data including 2022 (panel B).

By leveraging decomposition by EOFs, the spatiotemporal forecasting problem transitions into a univariate time series forecasting problem, as EOFs are inherently uncorrelated time series. This enables the utilization of various well–developed methodologies for time series forecasting. As a baseline, we employed classical autoregressive models available in the R package "modeltime"[22] (see Table 3). Additionally, we utilized Deep Learning autoregressive models, including DEEPAR, which is a DL architecture based on a Long Short–Term Memory (LSTM) Recurrent Neural Network[23], DEEP STATE, an approach to probabilistic time series forecasting that combines state space models with deep learning[24], NBEATS, a deep neural architecture based on backward and forward residual links and a very deep stack of fully–connected layers[25], and Gaussian Process (GP) Forecast, a DL architecture that automatically selects the optimal kernel in Gaussian process analysis of time series, while also providing reliable estimation of the hyperparameters[26], for forecasting purposes.

doi.org/10.32388/V71U3T

	Model	MAE	MAPE	MASE	SMAPE	RMSE
1	ARIMA	0.01	785.61	1.04	77.03	0.02
2	PROPHET	0.01	806.55	1.05	78.73	0.02
3	GLMNET	0.01	811.93	1.09	80.53	0.02
4	SVM–RBF	0.01	753.24	1.05	79.32	0.02
5	BOOST–TREE (H2O)	0.01	941.93	1.12	79.37	0.02
6	PROPHET–XGBOOST	0.02	918.38	1.18	83.03	0.02
7	RANDOM–FOREST	0.01	825.46	1.05	77.30	0.02

Table 3. Accuracy table for some autoregressive models.

While classic autoregressive models exhibited poor performance, DL models produced more accurate results. Table 4 summarizes the key metrics for comparing the models (**MAE**: Mean absolute error, **MAPE**: Mean absolute percentage error, **MASE**: Mean absolute scaled error, **SMAPE**: Symmetric mean absolute percentage error, and **RMSE**: Root mean squared error).

	Model	MAE	MAPE	MASE	SMAPE	RMSE
1	WAVELET–ANN[13]	0.015	2.942	2.349	0.019	0.971
2	DEEPAR[27]	0.016	833.	1.25	89.1	0.021
3	DEEP STATE[24]	0.014	661.	1.08	83.0	0.018
4	NBEATS[25]	0.014	813.	1.09	81.3	0.017
5	GAUSSIAN PROCESS FORECAST[26]	0.014	777.	1.04	77.7	0.017

Table 4. Accuracy table comparing different Deep Learning models in forecasting the EOF ϕ_1.

Fig. 5 depicts a comparison of forecasts for 2022 generated by the Wavelet–ANN Hybrid Model and the DEEPAR model using the first three EOFs. Precipitation records from three years between 2019 and 2021 were utilized as training data for model training.

In the remainder of this article, we will continue using the Wavelet–ANN Hybrid model for forecasting, which is built upon the wavelet transform using the Maximal Overlap Discrete Wavelet Transform (MODWT) algorithm

developed by Anjoy et al.[13].

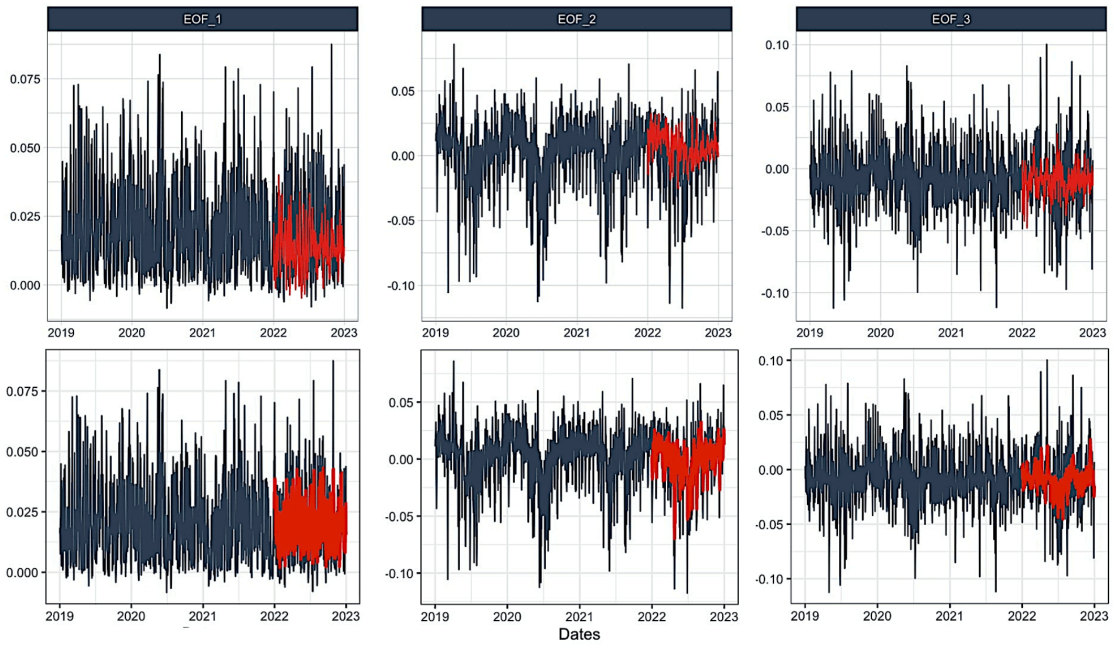

Figure 5. The first three EOF functions (in black) and their predictions (in red). The top three panels display predictions of the EOFs using a LSTM–Recurrent Neural Network with ten hidden layers, while the bottom three panels depict predictions using the Wavelet–ANN Hybrid Model.

For EOFs forescast with the Wavelet–ANN Hybrid model, we employed a Haar filter with 10 wavelet levels (the level of wavelet decomposition), and the size of the hidden layer = 40. Next, we reconstructed the spatiotemporal data for all locations within each cluster.

Fig. 6 offers a detailed illustration of the 2022 forecast using this method. By conducting the forecast separately for each cluster, we ensure similar precipitation patterns and more stable spatiotemporal relationships.

doi.org/10.32388/V71U3T

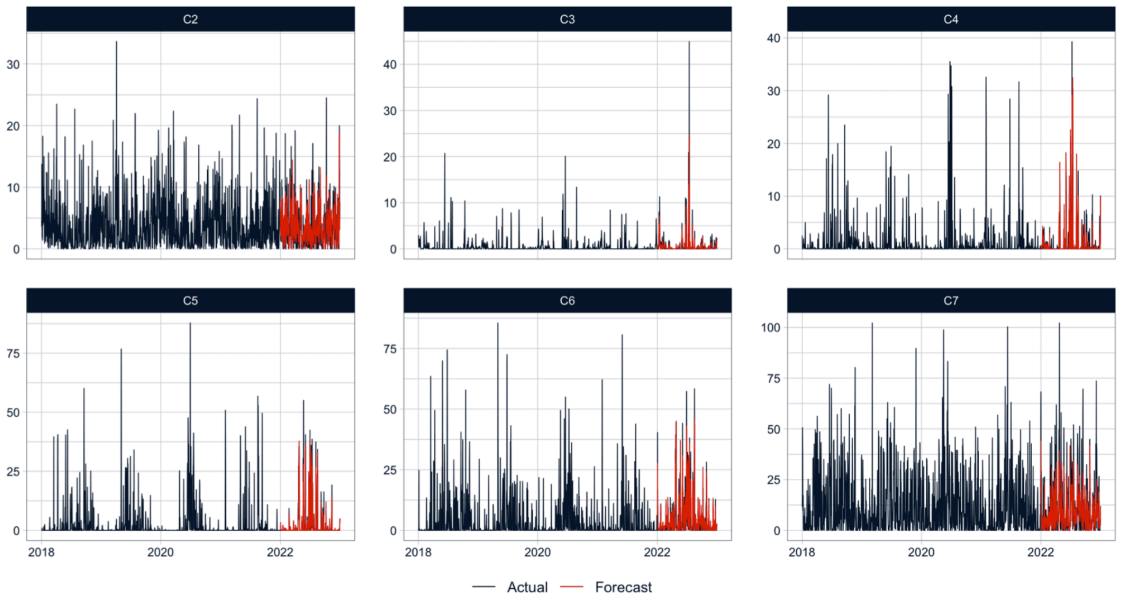

Figure 6. The precipitation time series for the medoids of clusters C2 to C7 for 2022 are shown. The black curves represent the actual precipitation records, while the red curves display the forecasts made using the EOF–Wavelet–ANN hybrid model.

We primarily use the medoid or representative of each cluster to validate the forecast. The medoid captures a characteristic precipitation pattern for each cluster and, consequently, for the geographic area it represents. The other points within this geographic area are expected to exhibit similar precipitation patterns. However, in larger clusters or those with poorer spatial coherence metrics, some time series within the cluster may deviate from the medoid's pattern. In these instances, we conduct multiple forecasts at various points within each cluster to thoroughly assess the model's performance.

Table 5 shows the forecast performance for each cluster medoid using the decomposition of spatiotemporal data into EOFs and capturing temporal patterns with Wavelet-ANN. The table shows the number of grid points (out of a total of 6355 points) within each cluster that were used to train the model.

To illustrate the complexity of applying the described model, the computation time for each data set corresponding to each cluster is indicated, using a MacBook Pro with a 2.3 GHz Intel Core i9 Eight-Core processor and 16 GB of 2667 MHz DDR4 memory. Predictions can be made for any grid point within the clusters, but for visualizing the predictions, the medoid of each cluster is used.

doi.org/10.32388/V71U3T

Cluster	Grid points	MAE	MASE	RMSE	Medoid(Long,Lat)	Expl.Variance(5EOFs)	Comput.Time
1	1232	1.423551	1.615929	2.122508	(-68.75,-20.25)	0.8148431(7EOFs)	39.18936 mins
2	1542	1.609527	0.4985094	2.280558	(-71.25,-53.5)	0.8998416	26.27438 mins
3	533	0.4189356	0.444972	1.348227	(-70.5,-29.25)	0.8478964	27.81493 mins
4	448	0.6333704	0.4218802	1.516656	(70.5,-33)	0.8785737	28.12317 mins
5	355	0.7353927	0.2167723	2.263974	(-71.5,-35.75)	0.8933317	27.90131 mins
6	580	2.022632	0.3941201	3.10687	(-72.25,-38.25)	0.8761673	27.94201 mins
7	1661	6.471836	0.6291881	10.05184	(-72.5,-44.25)	0.8931514	28.08024 mins

Table 5. Model Performance on the Precipitation Data for the Medoids of the Clusters.

Fig. 6 shows the precipitation time series for the medoids of clusters C2 to C7 for 2022. The locations of the medoids are indicated in table 5. The black curves represent the actual precipitation records, while the red curves show the forecasts made using the previously described method.

It can be observed that the forecasts appear reasonably accurate for these locations, with more precise results for the medoids of clusters C3 to C6. For the medoids of clusters C2 and C7, the forecasts seem less accurate, consistent with the indicators shown in Figure 6.

Time series for cluster C1 are not shown, but the results are less accurate. This is due to the low predictability indicated by the poor *DOF* and $var(SAI)$ metrics for these geographic areas.

4. Conclusions and Discussion

In this study, we presented an innovative approach to analyzing and forecasting spatiotemporal climatic patterns by integrating advanced statistical techniques with machine learning methods. By employing Empirical Orthogonal Functions (EOFs) for dimensionality reduction, we efficiently capture the essential temporal and spatial information within the climate data. The application of wavelet analysis provides high-resolution time-frequency information, enhancing the detail and accuracy of the forecasts. Machine learning methods are leveraged for their powerful predictive capabilities, allowing for the robust forecasting of a truncated number of EOFs over a medium-range horizon. Specifically the Wavelet-ANN Hybrid model, utilizing the Maximal Overlap Discrete Wavelet Transform (MODWT) algorithm, has proven to be effective in forecasting

the selected EOFs, thereby enabling the reconstruction of spatiotemporal data with extended temporal components.

This methodology not only addresses the complexities inherent in climate data analysis but also facilitates the transition from a high-dimensional multivariate forecasting problem to a more manageable low-dimensional univariate forecasting problem. This significant reduction in computational complexity is achieved without compromising the utility and accuracy of the forecasts.

Moreover, the use of cluster analysis with Dynamic Time Warping (DTW) for defining similarities between rainfall time series, along with spatial coherence and predictability assessments, has been instrumental in identifying geographic areas where model performance is enhanced. This approach also provides insights into the reasons behind poor forecast performance in regions or clusters with low spatial coherence and predictability. By using the medoids of the clusters, the forecasting process becomes more practical and efficient.

Overall, this study underscores the importance of combining statistical and machine learning techniques to tackle the intricate challenges posed by climate data analysis. The findings highlight the potential of this hybrid methodology to improve resource management and environmental planning, particularly in regions characterized by high climatic variability like Chile. The robust framework established in this research offers a pathway for more accurate and reliable climatic forecasts, ultimately contributing to better-informed decision-making processes in the face of climate change.

The primary motivation behind this study is to use pragmatic, computationally efficient models that provide useful forecasts while accounting for the climatic variability of an extensive and complex region such as Chile. Given this significant variability, we posit that a single predictive model or tool adaptable to all the intricate details scattered throughout the region is inherently unreliable.

To address the complexities of this problem, we employ a hybrid approach that combines several tools: (1) unsupervised classification (cluster analysis) to capture spatiotemporal correlations using Dynamic Time Warping (DTW)-based distances, (2) Empirical Orthogonal Function (EOF) decomposition on data classified (segmented) by clusters to reduce the problem's dimensionality, and (3) the application of machine learning methods on time series to capture temporal patterns in each geographic zone delineated by the clusters.

This approach yields localized forecasts for geographic regions exhibiting similar precipitation behavior patterns.

It is important to note that this method can also be applied to the entire dataset for a global forecast of the territory (i.e., without segmenting into geographic zones delineated by clusters) by increasing the number of EOFs (20 or more, to achieve more than 80% of the explained variance). However, this would result in reliable

forecasts in areas with better spatial coherence and predictability, but less reliable predictions in other zones where the data is not well represented. Segmenting the data into similarity patterns is a preliminary step to enhance the local nature of the forecast and can be beneficial for each specified geographic zone.

Furthermore, this approach can be combined with the innovative methodology recently proposed in[16], where spatial coefficients can be extended not only across the entire grid but also at any point within the territory using Deep Learning-based regression (interpolation). A dense Feed–Forward Neural Network (FFNN) captures local spatial patterns, facilitating the reconstruction of $\alpha(s)$ at any point. By passing the extended $\phi(t+h)$ over a horizon h to the final layer (recombination layer) of this FFNN for estimating spatial coefficients, the spatiotemporal fields are not only reconstructed at each grid point (potentially at any point, not limited to the grid) but also extended in time through the forecast of the associated time series.

References

1. ^*Seneviratne SI, et al. Climate Change 2018: The Physical Science Basis. Contribution of Working Group I to the Six th Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge, UK: Cambridge University Press; 2018.*
2. ^*Intergovernmental Panel on Climate Change (2014). Climate Change 2014: Impacts, Adaptation, and Vulnerabilit y. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Cha nge. Cambridge, UK: Cambridge University Press.*
3. ^*Cressie N, Wikle C. Statistics for Spatio-Temporal Data. Hoboken: Wiley; 2011.*
4. ^*Bochenek B, Ustrnul Z (2022). "Machine Learning in Weather Prediction and Climate Analyses: Applications and Perspectives". Atmosphere. **13** (2): 180. doi:10.3390/atmos13020180. Link to article.*
5. ^*Lam R, Sanchez-Gonzalez A, Willson M, Wirnsberger P, Fortunato M, Alet F, Ravuri S, Ewalds T, Eaton-Rosen Z, H u W, Merose A, Hoyer S, Holland G, Vinyals O, Stott J, Pritzel A, Mohamed S, Battaglia P (2023). "Learning skillful m edium-range global weather forecasting". Science. **382** (6677): 1416–1421. doi:10.1126/science.adi2336. PMID 37962 497.*
6. ^*Wong C (2023). "DeepMind AI accurately forecasts weather - on a desktop computer". Nature. Epub ahead of prin t. doi:10.1038/d41586-023-03552-y. PMID 37964116.*
7. ^*Cao D, Wang Y, Duan J, Zhang C, Zhu X, Huang C, Tong Y, Xu B, Bai J, Tong J, Zhang Q (2021). "Spectral Temporal G raph Neural Network for Multivariate Time-series Forecasting". arXiv. arXiv:2103.07719 [cs.LG].*
8. ^*Lorenz E. Empirical orthogonal functions and statistical weather prediction. Scientific report no. 1. Cambridge, M ass: Air Force Cambridge Research Center, Air Research and Development Command; 1956.*

9. ^a, ^b*Preisendorfer R. Principal Component Analysis in Meteorology and Oceanography. Amsterdam: Elsevier; 1988. (Developments in Atmospheric Science).*
10. ^*Monahan AH, Fyfe JC, Ambaum MHP, Stephenson DB, North GR (2009). "Empirical Orthogonal Functions: The M edium is the Message". Journal of Climate. **22** (24): 6501–6514. doi:10.1175/2009JCLI3062.1. Link to article.*
11. ^*Newman M, Sardeshmukh PD (1995). "A caveat concerning singular value decomposition". J. Clim.. **8**: 352--360.*
12. ^*Hannachi A, Joliffe I, Stephenson D (2007). "Empirical orthogonal functions and related techniques in atmospher ic science: A review". Int. J. Climatol.. **27**: 1119–1152. doi:10.1002/joc.1499.*
13. ^a, ^b, ^c, ^d*Anjoy P, Paul RK (2019). "Comparative performance of wavelet-based neural network approaches." Neural Comput and Applic. **31**: 3443–3453. doi:10.1007/s00521-017-3289-9.*
14. ^*Hersbach H, Bell B, Berrisford P, Biavati G, Hore1nyi A, Muf1oz Sabater J, Nicolas J, Peubey C, Radu R, Rozum I, Sch epers D, Simmons A, Soci C, Dee D, The9paut JN (2023). "ERA5 hourly data on single levels from 1940 to present." C opernicus Climate Change Service (C3S) Climate Data Store (CDS). doi:10.24381/cds.adbb2d47.*
15. ^*Jolliffe IT. Principal Component Analysis. 2nd ed. New York: Springer; 2002.*
16. ^a, ^b*Amato F, Guignard F, Robert S (2020). "A novel framework for spatio--temporal prediction of environmental da ta using deep learning". Sci Rep. **10**. doi:10.1038/s41598-020-79148-7.*
17. ^***CR2** (2024). Centro de Ciencia del Clima y la Resiliencia. Datos de Precipitaci☒3n. Accessed: 2024-05-29. Availabl e from: https://www.cr2.cl/datos-de-precipitacion/.*
18. ^*Giorgino T (2009). "Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package". Jour nal of Statistical Software. **31** (7): 1–24. doi:10.18637/jss.v031.i07.*
19. ^*Moron V, Robertson AW, Ward MN, Camberlin P (2007). "Spatial coherence of tropical rainfall at the regional scal e". Journal of Climate. **20**: 5244–5263.*
20. ^*Katz RW, Glantz MH (1986). "Anatomy of a rainfall index". Monthly Weather Review. **114**: 764–771.*
21. ^*Bretherton CS, Widmann M, Dynnikov VP, Wallace JM, Blade I (1999). "The effective number of spatial degrees of freedom of a time varying field". Journal of Climate. **12**: 1990–2009.*
22. ^*Dancho M (2024). modeltime.ensemble: Ensemble Algorithms for Time Series Forecasting with Modeltime. R pac kage version 1.0.3.9000, https://business-science.github.io/modeltime.ensemble/, https://github.com/business-scien ce/modeltime.ensemble.*
23. ^*Flunkert V, Salinas D, Gasthaus J (2017). "DeepAR: Probabilistic forecasting with autoregressive recurrent networ ks". CoRR. Available from: http://arxiv.org/abs/1704.04110.*
24. ^a, ^b*Rangapuram SS, Seeger MW, Gasthaus J, Stella L, Wang Y, Januschowski T (2018). "Deep State Space Models for Time Series Forecasting". Advances in Neural Information Processing Systems. **31**. Available from: https://proceedi ngs.neurips.cc/paper_files/paper/2018/file/5cf68969fb67aa6082363a6d4e6468e2-Paper.pdf.*

25. a, b *Oreshkin BN, Carpov D, Chapados N, Bengio Y (2019). "N-BEATS: Neural basis expansion analysis for interpretable time series forecasting". arXiv preprint arXiv:1905.10437. Available from: https://arxiv.org/abs/1905.10437.*

26. a, b *Salinas D, Bohlke-Schneider M, Callot L, Medico R, Gasthaus J (2019). "High-dimensional multivariate forecasting with low-rank Gaussian Copula Processes". Advances in Neural Information Processing Systems. Vol. **32**. Curran Associates, Inc. Available from: https://proceedings.neurips.cc/paper_files/paper/2019/file/0b105cf1504c4e241fcc6d519ea962fb-Paper.pdf.*

27. ^ *Salinas D, Flunkert V, Gasthaus J, Januschowski T (2020). "DeepAR: Probabilistic forecasting with autoregressive recurrent networks". International Journal of Forecasting. **36** (3): 1181–1191. doi:10.1016/j.ijforecast.2019.07.001. Link to article.*

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.