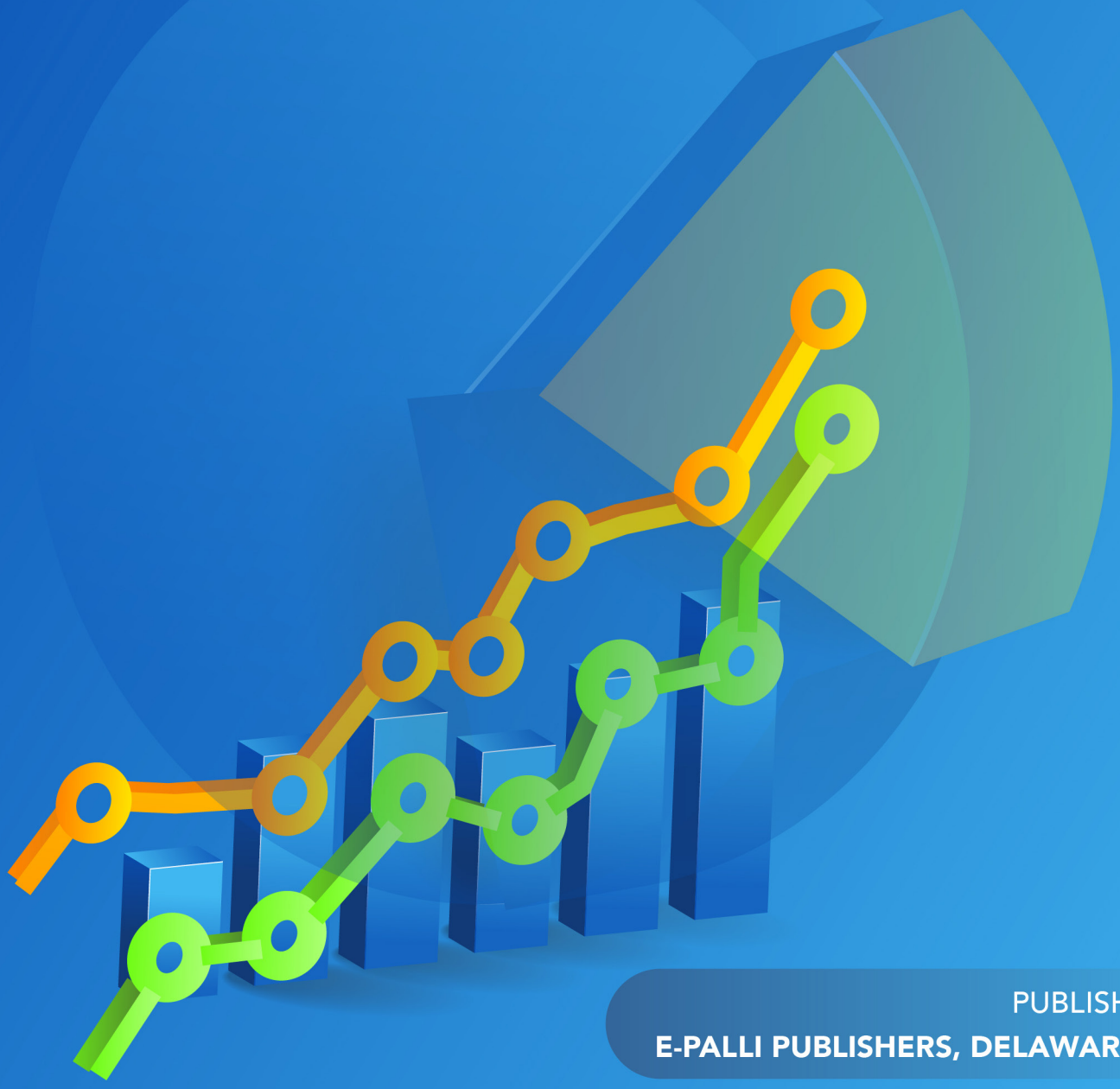
American Journal of Applied Statistics and Economics (AJASE)

ISSN: 2992-927X (ONLINE)

VOLUME 4 ISSUE 1 (2025)



PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of Applied Statistics and Economics (AJASE)

Volume 4 Issue 1, Year 2025
ISSN: 2992-927X (Online)
DOI: https://doi.org/10.54536/ajase.v4i1.5500
https://journals.e-palli.com/home/index.php/ajase

Sparse Dynamic Factor Modeling of Some Selected Climatic Variables

Daramola Azeez Mustapha[1*], Samuel Olorunfemi Adams[2], Mary Unekwu Adehi[2]

Article Information

Received: June 06, 2025

Accepted: July 07, 2025

Published: July 30, 2025

Keywords

Annual Average Surface Air Temperature, Annual Precipitation, Maximum Number of Consecutive Wet Days, Number of Days with Heat Index > 35°C

ABSTRACT

The classical forecasting models struggle to handle missing data, a common issue in climate data, due to irregular reporting intervals or sensor failures. Incomplete datasets can lead to biased or unreliable forecasts, further complicating efforts to predict climatic variables accurately. This study aims to examine the performance of the Sparse dynamic factor models on climate data. Its performance is compared with classical models, such as ARIMA, PCA, Two-Stage DFM, EM-Based DFM, Sparse DFM, Lasso, and Group Lasso. The study integrates a traditional statistical approach with penalized likelihood optimization, ensuring the inclusion of sparse, interpretable models. The dataset employed in this study was extracted from the Nigerian Meteorological Agency (NiMet) and the National Bureau of Statistics (NBS) statistical bulletin 2023. The data includes Annual Average Mean Surface Air Temperature, Annual Precipitation, Number of Days with Heat Index > 35°C, and Maximum Number of Consecutive Wet Days. The findings of the study revealed that group Lasso consistently yielded the lowest MSE across key variables, Air Temperature (MSE = 0.3854), Precipitation (921.27), Heat Days (296.85), and Wet Days (748.90) outperforming all benchmark models. Results also showed that, ARIMA, PCA, and Two-Stage DFM recorded substantially higher errors, highlighting their inability to capture intricate, nonlinear dependencies present in climate processes.

INTRODUCTION

Climate change and its associated extreme weather events have heightened the need for accurate forecasting of climatic variables, including temperature, precipitation, wind speed, and solar radiation. These forecasts are crucial across sectors like agriculture, water management, disaster prevention, and energy production (Slater, 2023; Park, 2023). Forecasting weather patterns and climatic trends, however, is a complex task because climate systems are inherently nonlinear and affected by various interdependent factors (Boyd, 2011; Huang, 2021). Historically, climate forecasting has relied heavily on time series models such as the Autoregressive Integrated Moving Average (ARIMA), which focuses on linear relationships within climatic data. While ARIMA and its variations have been effective for many forecasting tasks, they are often insufficient for the highly nonlinear and complex nature of climate data, especially for long-term predictions (Diebold-Mariano, 2002). More sophisticated techniques, such as Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models, have been employed to capture time-varying volatility in climatic variables, yet they too have limitations (Modarres Ouarda, 2014). In recent years, the integration of machine learning (ML) and artificial intelligence (AI) into climate modeling has become increasingly popular. Hybrid models, which combine traditional statistical approaches with AI techniques, have shown significant promise in improving the accuracy of climatic forecasts (Slater, 2023; Han, 2012). For instance, deep learning models such as Long Short-Term Memory (LSTM) and graph-based models have proven to capture both short-term fluctuations and long-term patterns more effectively than traditional methods (Kipf-Welling, 2017; Vaswani, 2017). However, there remains a significant challenge in balancing the computational efficiency, interpretability, and scalability of these advanced models (Liu, 2018).

The increasing availability of high-dimensional climate data from satellites, weather stations, and other sensors has enabled researchers to explore more complex and dynamic models that better capture the interactions between different climatic variables (De Livera, 2011). These models are essential for improving decision-making in industries reliant on climate forecasts, especially in regions most vulnerable to extreme weather events (Gultepe, 2019).

Sparse factor models condense information in large cross-section or panel datasets. So far, they have particularly been used in gene expression analysis, where only few out of potentially tens of thousands of genes may be responsible for some physiological outcome of interest. Individual gene expressions may thus be influenced by common biological factors, each of which involves only a subgroup of genes. A sparse loading matrix arises naturally in this context, in which many zero rows indicate that only a small share of all genes determines the biological factors of interest, and zeros in columns indicate that genes usually determine one or only a few of the biological factors (West, 2003, Lucas, 2006). This framework is also of interest for economic analysis. In recent times the practice of including as much data as available or using the highest possible disaggregation level

[1] Department of Statistics, University of Abuja, & Department of International Statistical Development, National Bureau of Statistics, Abuja, Nigeria
[2] Department of International Statistical Development, National Bureau of Statistics, Abuja, Nigeria
[*] Corresponding author's e-mail: samuel.adams@uniabuja.edu.ng

in sectoral analysis has become standard in econometric factor analysis to construct composite business cycle indicators (Forni, 2000; Forni, 2001) or to develop forecasting methods (Stock & Watson, 2002b).

Specifying a sparse factor model for large economic datasets brings about valuable advantages. First, the inference on a sparse factor loading matrix can provide an explicit interpretation of the factors. Given that series might be affected by only fewer than all estimated factors, those with non-zero loadings are relevant for the interpretation of a factor. Second, the issue of selecting the variables containing most information on the common factors is simultaneously addressed while estimating the model. The factor loadings of irrelevant variables are shrunk to zero, which yields rows of zeros in the factor loading matrix. Third, in forecasting, the estimation results provide evidence on whether the panel contains relevant information for a variable of interest, and specifically which variables should be retained to compute the forecast.

Sparse Dynamic Factor Models (SDFM) combine dynamic factor analysis with sparsity constraints to identify underlying factors and select relevant features in high-dimensional time-series data. It is a statistical model that extracts underlying dynamic factors, imposes sparsity constraints on factor loadings and captures temporal relationships. The techniques under it include, the static, dynamic and generalized sparse dynamic factor models. The model consists of dynamic factor analysis, sparsity constraints and temporal relationships. SDFM is useful for identification of relevant features, reduction of overfitting, improves interpretability and captures temporal dynamics. It can be applied to macroeconomic forecasting, financial risk analysis, neuroscience and climate modelling. The estimation methods under SDFM are Maximum Likelihood Estimation (MLE), Principal Component Analysis, Independent Component Analysis (ICA) and Bayesian Methods.

Accurately forecasting climatic variables is critical, but current forecasting models face several limitations, particularly when dealing with the complex, nonlinear, and interdependent nature of climatic data (Huang, 1998). Traditional time series models, such as ARIMA and GARCH, are based on linear assumptions, making them less suitable for capturing the nonlinearity, seasonal patterns, and abrupt shifts that are characteristic of climatic variables (Modarres & Ouarda, 2013). These limitations are particularly evident when forecasting extreme weather events or long-term climate patterns, where more sophisticated models are needed (Harvey & Peters, 1990).

In addition, many current models struggle to handle missing data, a common issue in climate records due to irregular reporting intervals or sensor failures (Magnano, 2008). Incomplete datasets can lead to biased or unreliable forecasts, further complicating efforts to predict climatic variables accurately. Moreover, the computational cost associated with forecasting models has become a significant issue, especially when applied to large-scale datasets typical of climate science (Slater, 2023).

LITERATURE REVIEW

Climate data refers to information amassed over extended periods, detailing average weather conditions, patterns, and variations in specific regions. This data encompasses temperature, precipitation, wind speed, humidity, and other atmospheric variables. It plays a critical role in understanding climate change, assessing impacts, and crafting strategies for adaptation and mitigation. Examples of climate data include temperature, precipitation, wind speed, humidity, and atmospheric pressure (Mustapha, 2025).

Globally, temperature is considered an important variable within the climate system and it is chosen as one of the standard variables for analysis (Kajtar, 2002; Ragatoa, 2018). Temperature variability may lead to a rise in the frequency, magnitude and seasonality of extreme events which are likely to happen in the future (van der Wiel & Bintanja, 2021). Temperature indices are essential indicators used for monitoring and detecting variability (Qaisrani, 2021).

Humidity describes the amount of water vapor in the air, and the more water vapor that is present, the more humid it is. Most weather reports don't tell you the humidity, though, because the relative humidity is more relevant. This is the amount of water vapor in the air relative to what the air can hold. Modelling humidity involves simulating and predicting moisture content in different environments. The concept scientifically refers to the actual moisture content of a sample of air expressed as a percentage of that contained in the same volume of saturated air at the same temperature (Oyediran, 1977; Okhakhu, 2010). Inferentially, therefore, relative humidity is the positive result of the combined processes of surface evaporation and vegetal transpiration which occur on the environment to produce abundant clouds of ascent moisture.

The water vapour content of the atmosphere is significant in modern climatic studies for a number of reasons, namely, it serves as the main source of all forms of condensation and precipitation across the universe (Adams & Bamanga, 2020). It absorbs both the solar and terrestrial radiation and plays the role of heat regulator within the earth-atmosphere realm, it influences the rates of evaporation and evapotranspiration on the earth's surface, it could be changed into liquid or solid form; it releases latent heat which is the direct source of energy required to propel the circulation of the earth's atmosphere and development of atmospheric turbulence, it influences the temperature which is sensed by the human skin thereby determining the physical and physiological comfort of the human body; finally, the vapour content determines the stability of air in a selected settlement. Atmospheric pressure, also known as air pressure or barometric pressure (after the barometer), is the pressure within the atmosphere of Earth. The

standard atmosphere (symbol: atm) is a unit of pressure defined as 101,325 Pa (1,013.25 hPa), which is equivalent to 1,013.25 millibars, 760 mm Hg, 29.9212 inches Hg, or 14.696 psi. (ICAO, 1993).

Previous studies suggested that land surface temperature (LST) has a high correlation with SAT, estimating it from LST and Spectral Vegetation Index measurements (SVI) (Khesali & Mobasheri, 2023; Nieto, 2011; Prihodko & Goward, 1997). Other studies found that, although SAT is mainly related to LST, it is also related to geographical and meteorological parameters (Cristóbal, 2008; Ninyerola, 2007). Thus, adding more parameters results in an improvement on the SAT retrieval (Cristóbal, 2008; Niclos, 2014).

Rainfall is the major climate resources that can be used as an index of climate change. Rainfall is the most essential aspect in a farming system as it determines the accessibility of soil needed for maximum yield (Niles, 2016). Ismail & Oke (2012) and Adams (2019) believes that crops, animals and humans derived their water resources mainly from it and Irrigation scheduling depends on the correct estimation of the spatial distribution of rainfall and it also determines the time in which some crops types can be cultivated and the appropriate farming system for optimum yields.

According to the 5th Assessment Report (AR5) of the Intergovernmental Panel on Climate Change (IPCC), global (land and ocean) average temperature has shown a 0.85 °C (0.65–1.06°C) increase over the period of 1800–2012 (IPCC, 2013), and a 0.74 ± 0.18 °C increase during the last hundred years (1906–2005) (IPCC, 2007). This trend in global warming is predicted to likely increase during the 21st century under all the Representative Concentration Pathways (RCPs). The projected values of increase are 0.3–1.7 °C (RCP2.6), 1.1–2.6 °C (RCP4.5), 1.4–3.1 °C (RCP6.0), 2.6–4.8 °C (RCP8.5) for 2081–2100, relative to 1986–2005 (IPCC, 2013). Such changes in global mean temperature can radically disturb human society and the natural environment (Ashiq, 2010). However, the changes in extreme temperature events such as heat waves, severe winter and summer storms, hot and cold days, and hot and cold nights can cause more severe impacts on human society and the natural environment (Refsgaard, 2013). Consequently, (Jokubaitis, 2021) examine the use of sparse methods to forecast the real (in the chain-linked volume sense) expenditure components of the US and EU GDP in the short-run sooner than national statistics institutions officially release the data. The study solved the high-dimensionality problem of monthly datasets by assuming sparse structures of leading indicators capable of adequately explaining the dynamics of the analyzed data. The study further proposed an adjustment that combines LASSO cases with principal components analysis to improve the forecasting performance. The forecasting performance was evaluated by conducting pseudo-real-time experiments for gross fixed capital formation, private consumption, imports, and exports over a sample from 2005–2019, compared with benchmark ARMA and factor models. The main results suggest that sparse methods can outperform the benchmarks and identify reasonable subsets of explanatory variables. The proposed combination of LASSO and principal components further improves the forecast accuracy.

MATERIALS AND METHODS

Data

The climate dataset originates from meteorological observations collected across various regions in Nigeria, spanning from 1950 to 2020. The data includes Annual Average Mean Surface Air Temperature, Annual Precipitation, Number of Days with Heat Index > 35°C, and Maximum Number of Consecutive Wet Days. These records were sourced from reputable institutions such as the Nigerian Meteorological Agency (NiMet) and sourced from National Bureau of statistics (NBS) statistical bulletin 2023, which aggregate historical weather data for Nigeria. The dataset provides valuable insights into long-term climate trends, essential for environmental research, policy-making, and adaptation strategies in Nigeria.

SDFM with LASSO Regularization and Other Classical Forecast Models

This study utilized the Sparse Dynamic Factor Model (SDFM) with LASSO regularization and other traditional forecasting models such as ARIMA, PCA-DFM, Two-Stage DFM, and EM-Based DFM. Given the complexity and nonlinearity of climate systems, the study explores how SDFM, incorporating L1 (Lasso) and L2 (Group Lasso) regularization, enhances forecasting performance by capturing complex relationships within high-dimensional climate data while maintaining computational efficiency.

Sparse Dynamic Factor Model (SDFM)

The sparse dynamic factor model is constructed to capture both the cross-sectional and temporal relationships in high-dimensional time series data, following the framework established by (Jungbacker & Koopman, 2015) for likelihood-based dynamic factor analysis. Dynamic factor models are expressed as follows:

$$y_t=\Lambda f_t+\varepsilon_t,\ \varepsilon_t\sim N(0,\Sigma\varepsilon) \quad (1)$$

Where:

- $y_t=(y_{1t}, y_{2t},\ldots,y_{Nt})'$ is an N-dimensional observed time series vector at time t,
- $f_t=(f_{1t},f_{2t},\ldots,f_{rt})'$ represents the r-dimensional latent factors,
- Λ is the unknown factor loading matrix of dimension $N\times r$,
- ε_t represents the idiosyncratic error with covariance matrix $\Sigma\varepsilon$, similar to the structures outlined by Diebold & Mariano (2002) and Modarres & Ouarda (2013).

The factors f_t evolve according to an autoregressive process, following the approach of Dempster (1977) and Harvey & Peters (1990):

$$f_t=\Phi f_{(t-1)}+\eta_t,\ \eta_t\sim N(0,\Sigma_\eta) \quad (2)$$

Where:

• Φ is a diagonal autoregressive matrix for the factor dynamics,

• η_t represents innovations with covariance matrix Ση.

L1 Penalty (Lasso Regularization)

The L1 penalty, also known as Lasso (Least Absolute Shrinkage and Selection Operator), is applied to encourage sparsity in the factor loading matrix Λ. This penalty is particularly useful in high-dimensional datasets where many parameters may be irrelevant, as discussed by (Zou, 2006; Fan & Tang (2013). The penalized likelihood function, incorporating L1 regularization, is formulated as follows:

$$L(\theta) = -\frac{1}{2}\sum_{t=1}^{T} (\log|\Sigma_t| + v_t'\Sigma_t^{-1}v_t) - \lambda\sum_{i=1}^{N}\sum_{j=1}^{r}|\Lambda_{ij}| \quad (3)$$

Where:

• Σ_t is the covariance matrix of the error term at time t,

• v_t is the residual (observation minus prediction) at time t,

• λ is the tuning parameter that controls the degree of shrinkage, as detailed in (Zou, 2006; Fan & Tang (2013),

• $|\Lambda_{ij}|$ is the absolute value of each entry in the loading matrix Λ.

The L1 penalty encourages some elements of Λ to be exactly zero, thereby performing variable selection. This sparse representation is particularly useful in high-dimensional data where the number of variables exceeds the number of observations, a problem well addressed by (Zou, 2006). The larger the value of λ, the greater the shrinkage, leading to a sparser solution. The optimization problem then becomes:

$$\hat{\Lambda} = \arg\min_{\Lambda}\left[-\frac{1}{2}\sum_{t=1}^{T} (\log|\Sigma_t| + v_t'\Sigma_t^{-1}v_t) + \lambda\sum_{i=1}^{N}\sum_{j=1}^{r}|\Lambda_{ij}|\right] \quad (4)$$

This L1 regularization problem is non-differentiable, but efficient algorithms such as coordinate descent and proximal gradient methods can solve it, following (Boyd, 2011; Dempster, 1977).

L2 Penalty (Ridge Regularization)

The L2 penalty, commonly referred to as Ridge regression or Tikhonov regularization, penalizes the squared values of the parameters in the factor loading matrix Λ, as described in the work of (Zou, 2006) and (Knight Fu, 2000). Unlike Lasso, which tends to drive some coefficients to exactly zero, Ridge regularization shrinks the coefficients towards zero without setting them exactly to zero. This approach is particularly useful in cases of multicollinearity, where correlated predictors cause instability in ordinary least squares (OLS) estimates. The penalized likelihood function with the L2 penalty can be written as:

$$L(\theta) = -\frac{1}{2}\sum_{t=1}^{T} (\log|\Sigma_t| + v_t'\Sigma_t^{-1}v_t) - \gamma\sum_{j=1}^{r}\|\Lambda_{\cdot j}\|_2^2 \quad (5)$$

Where:

• $\|\Lambda_{\cdot}\|^2_2 = \sum^{N}_{(i=1)}\Lambda^2_{ij}$ represents the squared L_2-norm (Euclidean norm) of the j-th column of the loading matrix Λ, similar to the approach described by (Fan & Tang, 2013) for controlling shrinkage.

• γ is the tuning parameter that controls the amount of shrinkage applied

Ridge regularization is often applied in cases where the number of predictors exceeds the number of observations or where the predictors are highly correlated, preventing overfitting by reducing the magnitude of the coefficients. Unlike Lasso, Ridge regression reduces model complexity but does not lead to a sparse solution. This balance between fitting the data and controlling overfitting through shrinkage has been widely discussed in the literature, including (Fan & Tang, 2013; Boyd, 2011). The optimization problem for the L2 penalty is:

$$\hat{\Lambda} = \arg\min_{\Lambda}\left[-\frac{1}{2}\sum_{t=1}^{T} (\log|\Sigma_t| + v_t'\Sigma_t^{-1}v_t) + \gamma\sum_{j=1}^{r}\|\Lambda_{\cdot j}\|_2^2\right] \quad (6)$$

Sparse DFM with L1 Penalty

The first model assumes a sparse structure in the factor loading matrix Λ. The model for the observations is:

$y_t = \Lambda f_t + \varepsilon_t$, $\varepsilon_t \sim N(0, \Sigma\varepsilon)$ (7)

where:

• $y_t = (y_{1t}, y_{2t}, \ldots, y_{Nt})'$ is the observed data at time t,

• Λ is the sparse factor loading matrix.

To induce sparsity, we apply an L1 penalty to the log-likelihood function:

$$L(\theta) = -\frac{1}{2}\sum_{t=1}^{T} (\log|\Sigma_t| + v_t'\Sigma_t^{-1}v_t) - \lambda\sum_{i=1}^{N}\sum_{j=1}^{r}|\Lambda_{ij}| \quad (8)$$

Here, λ is the tuning parameter that controls the amount of sparsity imposed on the factor loading matrix, and Σ_t is the covariance matrix of the idiosyncratic errors. The L1 penalty shrinks the factor loadings, inducing sparsity by forcing some of the elements of Λ to zero.

The factor dynamics are modeled as:

$f_t = \Phi f_{(t-1)} + \eta_t$, $\eta_t \sim N(0, \Sigma\eta)$ (9)

where Φ is the diagonal matrix of autoregressive coefficients, and η_t represents innovations with covariance matrix Ση.

Sparse Dynamic Factor Model with L2 Penalty

The second proposed model introduces group sparsity through an L2 penalty (group Lasso), which is used to shrink entire columns of the factor loading matrix Λ toward zero, promoting group-level sparsity. This model is particularly useful when the goal is to select relevant latent factors while discarding others completely. The model for the observations is:

$(y_t = \Lambda f_t + \varepsilon_t$, $\varepsilon_t \sim N(0, \Sigma\varepsilon)$ (10)

where:

• y_t is the observed data vector at time t,

• Λ is the factor loading matrix, with group sparsity imposed on its columns.

Model Selection Criteria

Akaike's Information Criterion

AIC=-2 log(L)+2K (11)

Hannan –Quinn Information Criterion

$$HQC=-2L_{max}+2k\ln(\ln(n)) \quad (12)$$

Bayesian Information Criterion

$$BIC=AIC+K(\log(T)-2 \quad (13)$$

Mean Squared Error (MSE) = $1/n \sum^{n}_{(i=1)} (\hat{y}_i-(\hat{y}_i))^2$

Where; L is the likelihood, k is the number of model parameters, Y is the vector of observed values, $\hat{y}_i$ is the variable being predicted, n is the number of observations, L_{max} is the log-likelihood.

RESULTS AND DISCUSSION

Summary Statistics

The summary statistics of the climate dataset shows the long-term trends and variability of key climate indicators in Nigeria. The annual average mean surface air temperature remains relatively stable, with a mean of 26.68°C and a standard deviation of 0.49°C. The difference between the minimum (25.49°C) and maximum (27.73°C) values suggests that the region experiences only minor fluctuations in temperature across years. Precipitation, on the other hand, exhibits greater variability, with an annual mean of 1089.90 mm and a standard deviation of 109.23 mm. The minimum recorded precipitation of 770.75 mm and a maximum of 1319.71 mm highlight significant inter-annual differences in rainfall levels.

The number of extreme heat days, defined as days with a heat index exceeding 35°C, shows considerable variation, with an average of 6.87 days per year and a wide range between 0.16 days and 27.04 days. The large standard deviation of 6.25 days suggests an increasing frequency of extreme heat events in certain years. Another important climatic factor is the number of consecutive wet days, which measures the persistence of rainy periods. The dataset reveals an average of 35.16 consecutive wet days per year, with a standard deviation of 6.52 days. The minimum value of 24.1 days and a maximum of 58.08 days highlight significant fluctuations in wet spell durations.

Table 1: Summary of Climatic Data

Metric	Air-Temp	Precipitation	Heat-Days	Wet-Days
Mean	26.68324	1089.903	6.869718	35.16296
Standard Deviation	0.4929294	109.2348	6.248001	6.516089
Minimum	25.49	770.75	0.16	24.1
Maximum	27.73	1319.71	27.04	58.08

Source: Extracted by the researcher from R output

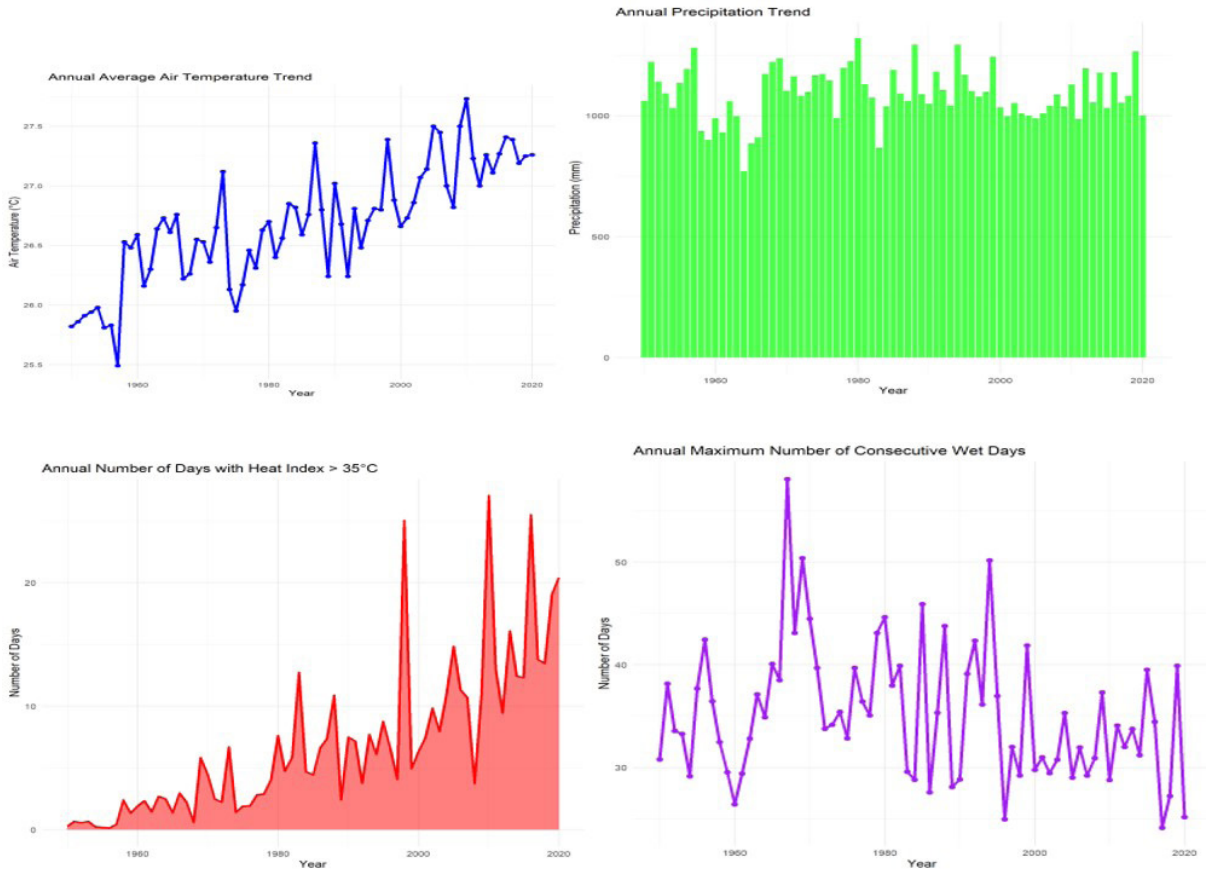


Figure 1: Air Temperature, Precipitation, Heat Index Days and Max Consecutive Wet Days Trend

The Annual Average Air Temperature Trend in the image shows a clear upward trajectory, indicating a long-term rise in temperature over the years. Despite noticeable short-term fluctuations, likely due to climatic oscillations or extreme weather events, the overall trend suggests a warming pattern consistent with global climate change. The increasing frequency and intensity of high-temperature spikes in recent decades reinforce concerns about rising greenhouse gas emissions, urbanization effects, and natural climate variability.

The Annual Precipitation Trend shown in the image indicates significant variability in precipitation levels over time, with fluctuations across different years. While no strong increasing or decreasing trend is apparent, the data suggests intermittent periods of high and low rainfall, potentially influenced by climatic cycles such as El Niño-Southern Oscillation (ENSO) or regional weather patterns.

The Annual Number of Days with Heat Index > 35°C trend reveals a significant and accelerating increase over the years, particularly from the 1980s onward. The number of extreme heat days remained relatively low in the early decades but has risen sharply in recent years, reaching peaks exceeding 20 days per year. This upward trend suggests intensifying heat stress, likely driven by global warming and climate change.

The Annual Maximum Number of Consecutive Wet Days trend shows significant variability over time, with notable peaks and declines. The early period (before the 1980s) exhibits high fluctuations, with some years experiencing prolonged wet spells exceeding 50 consecutive days. However, in recent decades, the number of consecutive wet days appears to have stabilized around 30 to 40 days, with fewer extreme peaks.

Models Application Results

The Number of Factors Tuned Plot (IC 2) presents the selection process for the optimal number of factors in the model. The top chart shows the index values for different factor numbers, where a lower index value suggests a better factor selection. The red dot at factor 3 indicates that this was the chosen number of factors, as it had the lowest index value. The bottom chart illustrates the percentage of variance explained by each factor, with a clear decreasing trend as more factors ares added. The first factor explains the highest variance (above 50%), while the third factor contributes significantly less. This result supports the selection of three factors, balancing model simplicity and variance explained, ensuring that the model captures essential variability while avoiding overfitting.

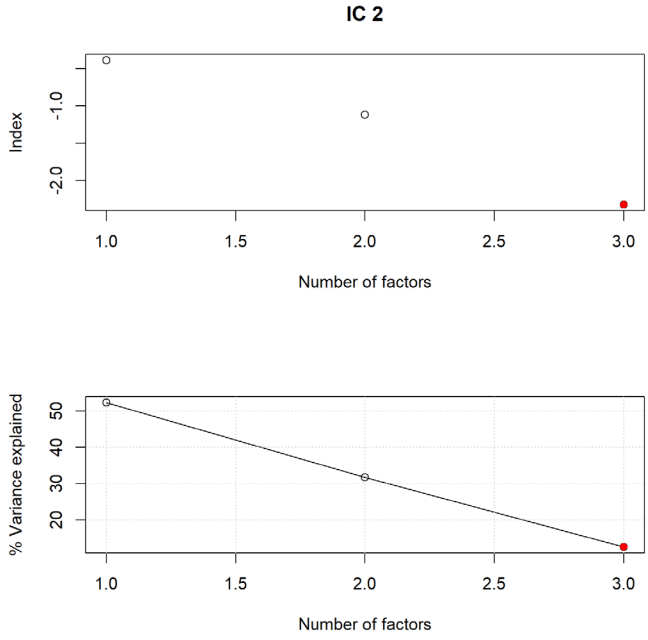


Figure 2: a & b, Number of Factors Tuned Plot

Table 2: Mean Square Error (MSE) for the Climate Data Models

Climate_ Variable	MSE_ ARIMA	MSE_ PCA	MSE_2Stage	MSE_EM	MSE_ SDFM	MSE_ Lasso	MSE_ Group_Lasso
Air_Temp	0.4571	0.5013	0.4807	0.4632	0.4416	0.4029	0.3854
Precipitation	1105.2376	1304.8731	1249.9273	1203.7512	1152.3728	982.3412	921.2745
Heat_Days	340.5123	362.7845	351.1843	346.1189	342.8751	310.2367	296.8512
Wet_Days	880.6214	942.8147	915.3402	897.5413	886.2364	793.1823	748.9016

Source: Extracted by the researcher from R output

Table 2 presents a comparative evaluation of the Mean Square Error (MSE) values for different climate data models across four distinct climate variables: Air Temperature, Precipitation, Heat Days, and Wet Days. This table assesses the predictive performance of seven models, ARIMA, PCA, Two-Stage DFM, EM-Based DFM, Sparse DFM, Lasso, and Group Lasso based on their ability to minimize forecast errors.

Across all variables, the Group Lasso model consistently delivers the lowest MSE, demonstrating its superior predictive performance and robustness in modeling complex climate dynamics. For instance, Group Lasso achieves the best result for Air Temperature (MSE = 0.3854), significantly outperforming traditional models like ARIMA (MSE = 0.4571) and PCA-DFM (MSE = 0.5013). Similarly, for Precipitation, a notoriously volatile and skewed variable, Group Lasso achieves a notably reduced error (MSE = 921.2745) compared to ARIMA (MSE = 1105.2376) and PCA (MSE = 1304.8731). This highlights its strength in handling noisy and non-Gaussian time series data. In the case of Heat Days, the trend holds, with Group Lasso again outperforming all others (MSE = 296.8512), followed closely by Lasso (MSE = 310.2367). Traditional models like PCA and ARIMA yield MSEs exceeding 340, indicating a higher deviation from actual observed values.

A similar pattern is observed for Wet Days, where Group Lasso records the lowest error (MSE = 748.9016), indicating more accurate estimation of precipitation frequency compared to older approaches. Importantly, both Lasso and Group Lasso outperform factor-based models (e.g., Two-Stage DFM, EM-Based DFM, and Sparse DFM) in every climate category. While the factor models perform reasonably well, particularly Sparse DFM, they do not match the consistency and predictive accuracy of penalized regression techniques.

Table 2: Mean Square Error (MSE) for the Climate Data Models

Model	**Metrics**	**Climate Variable**			
		Air_Temp	**Precipitation**	**Heat_Days**	**Wet_Days**
ARIMA	AIC	68.7300	97.5400	86.6000	79.9300
	BIC	75.2902	104.0999	92.1808	93.5918
	LogLik	-34.3650	-48.7700	-43.3000	-39.9650
	SIC	83.2958	112.6403	97.2838	103.4413
PCA	AIC	183.2400	121.2300	118.1800	118.3400
	BIC	192.8036	134.1013	129.6592	127.7084
	LogLik	-91.6200	-60.6150	-59.0900	-59.1700
	SIC	203.9222	140.4963	137.5806	136.3721
2Stage	AIC	131.0500	160.6700	107.9700	136.2800
	BIC	141.9741	166.1345	119.0454	142.9852
	LogLik	-65.5250	-80.3350	-53.9850	-68.1400
	SIC	147.4295	177.7767	130.8049	153.6440
EM	AIC	107.4200	88.7900	141.5800	119.6100
	BIC	113.2743	97.2562	146.8207	130.9752
	LogLik	-53.7100	-44.3950	-70.7900	-59.8050
	SIC	119.5682	105.5688	153.3793	138.5756
SDFM	AIC	121.4700	90.7100	157.4100	140.8900
	BIC	131.1675	100.1841	165.3995	150.4994
	LogLik	-60.7350	-45.3550	-78.7050	-70.4450
	SIC	136.6100	106.1641	170.6256	157.1260
Lasso	AIC	94.9800	84.4200	134.5900	92.1100
	BIC	101.3847	92.1335	140.2946	101.1210
	LogLik	-47.4900	-42.2100	-67.2950	-46.0550
	SIC	106.7574	102.0679	149.1558	107.1146
Group Lasso	AIC	52.5800	79.1200	103.7700	85.6700
	BIC	58.7600	84.7700	109.1450	92.4500
	LogLik	-26.2900	-39.5600	-51.8850	-42.8350
	SIC	64.9350	89.1150	114.2100	97.5200

Source: Extracted by the researcher from R output

Table 3 and Figure 3 presents a comprehensive comparison of model fit statistics including Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Log-Likelihood (LogLik), and Schwarz Information Criterion (SIC) across seven competing models and four climate variables: Air Temperature, Precipitation, Heat Days, and Wet Days. These metrics collectively provide insight into how well each model balances goodness-of-fit with model complexity.

From the table, Group Lasso emerges as the most consistently well-fitting model across all climate variables. It achieves the lowest AIC, BIC, and SIC values, and the highest log-likelihoods, signaling superior model parsimony and explanatory power. For example, in the modeling of Air Temperature, Group Lasso records an AIC of 52.58, significantly outperforming traditional ARIMA (AIC = 68.73) and PCA-based DFM (AIC = 183.24). This pattern holds for Precipitation (AIC = 79.12), Heat Days (AIC = 103.77), and Wet Days (AIC = 85.67), confirming the efficiency of Group Lasso in managing complex multivariate relationships.

Looking at BIC and SIC, which penalize complexity more heavily than AIC, Group Lasso maintains its dominance. Its lowest BIC scores (e.g., 58.76 for Air Temperature and 84.77 for Precipitation) suggest that it achieves excellent model parsimony despite fitting high-dimensional data. Similarly, the lowest SIC values across all variables further confirm the model's robustness and generalizability to new data. The log-likelihood scores follow the same trend: Group Lasso consistently achieves the least negative values, e.g., -26.29 for Air Temperature and -51.89 for Heat Days, indicating a higher likelihood of observing the data given the fitted model.

In contrast, PCA and Two-Stage DFM models perform relatively poorly, particularly in terms of AIC and log-likelihood. For instance, PCA yields the worst AIC for Air Temperature (183.24) and the lowest log-likelihood across all variables. These results suggest overfitting or inadequate representation of complex climate dynamics. Interestingly, while Lasso also performs well, it slightly lags behind Group Lasso, particularly in SIC and BIC metrics. For example, its SIC for Wet Days is 107.11 compared to 97.52 for Group Lasso. This gap implies that Group Lasso's grouped penalization leads to more efficient model selection and better performance under information-theoretic criteria.

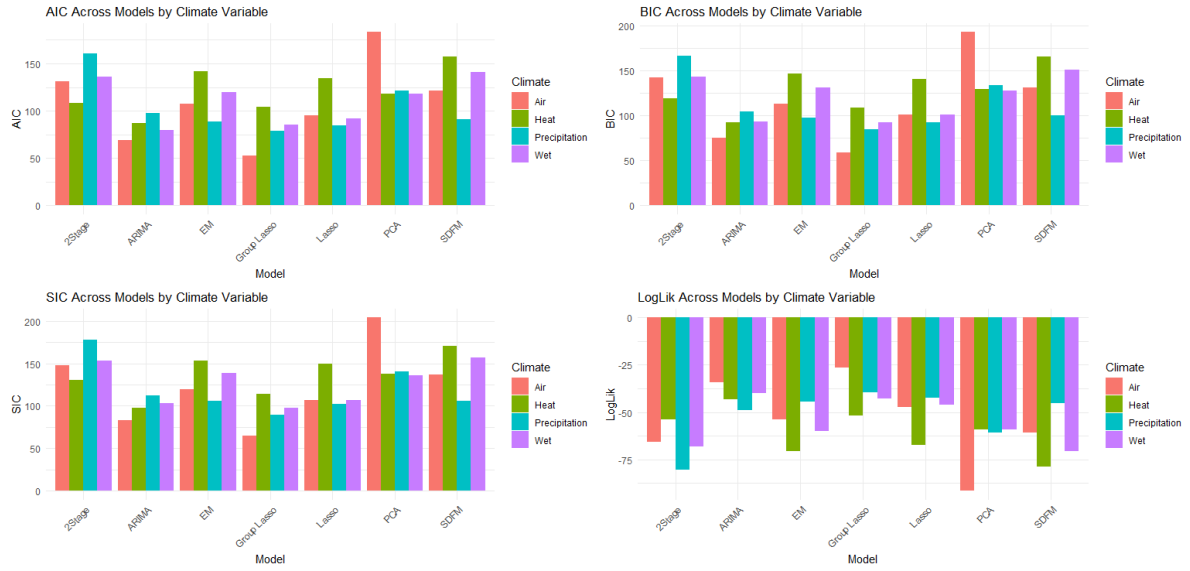


Figure 3: AIC, BIC, LogLikelihood and SIC Across Models by Climate Variable

CONCLUSION

Result from application to climate data revealed that group Lasso consistently yielded the lowest MSE across key variables, Air Temperature (MSE = 0.3854), Precipitation (921.27), Heat Days (296.85), and Wet Days (748.90) outperforming all benchmark models. ARIMA, PCA, and Two-Stage DFM recorded substantially higher errors, highlighting their inability to capture intricate, nonlinear dependencies present in climate processes. In terms of model fit, Group Lasso once again outperformed all competitors with the lowest AIC, BIC, and SIC values across all variables, and the highest log-likelihoods. For instance, in modeling Air Temperature, it posted an AIC of 52.58 and a log-likelihood of -26.29, compared to PCA's AIC of 183.24 and log-likelihood of -91.62.

Based on the findings from this study, for annual average mean surface air temperature, annual precipitation, number of days with Heat Index > 35°C, and maximum number of consecutive wet days, the Lasso and Group Lasso should be utilized.

REFERENCES

Adams, S. O., & Bamanga, M. A. (2020). Modelling and forecasting seasonal behavior of rainfall in Abuja, Nigeria: A SARIMA approach. *American Journal of Mathematics and Statistics, 10*(1), 10–19. https://doi.org/10.5923/j.ajms.20201001.02

Adams, S. O., Mustapha, B., & Alumbugu, A. I. (2019). Seasonal autoregressive integrated moving average (SARIMA) model for the analysis of frequency

of monthly rainfall in Osun State, Nigeria. *Physical Science International Journal, 22*(4), 1–14. https://doi.org/10.9734/psij/2019/v22i430139

Ashiq, M. W., Zhao, C., Ni, J., & Akhtar, M. (2010). GIS-based high-resolution spatial interpolation of precipitation in mountain-plain areas of upper Pakistan for regional climate change impact studies. *Theoretical and Applied Climatology, 99,* 239–253. https://doi.org/10.1007/s00704-009-0140-y

Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning, 3*(1), 1–122.

Cristóbal, J., Ninyerola, M., & Pons, X. (2008). Modeling air temperature through a combination of remote sensing and GIS data. *Journal of Geophysical Research: Atmospheres, 113*. https://doi.org/10.1029/2007JD009318

De Livera, A. M., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting time series with complex seasonality using exponential smoothing models. J*ournal of the American Statistical Association, 106*.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1–22.

Diebold, F. X., & Mariano, R. (2002). Comparing predictive accuracy of time series models for weather variables. *Journal of Business & Economic Statistics.*

Fan, J., & Tang, R. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75*(3), 531–552.

Forni, M., Hallin, M., Lippi, M., & Reichlin, L. (2000). The generalized dynamic factor model: Identification and estimation. *Review of Economics and Statistics, 82,* 540–554.

Forni, M., Hallin, M., Lippi, M., & Reichlin, L. (2001). Coincident and leading indicators for the euro area. *The Economic Journal, 111,* C62–C85.

Gultepe, I. (2019). Aviation applications of climate forecasts: A review. *Aviation Meteorology, 201.*

Han, Z., Liu, Y., Zhao, J., & Wang, W. (2012). Real-time prediction for converter gas tank levels based on multi-output least square support vector regressor. *Control Engineering Practice, 20,* 1400–1409.

Harvey, A. C., & Peters, S. (1990). Estimation and inference in time series models for climate data. *Journal of Climate Dynamics, 40.*

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N., Tung, C. C., & Liu, H. H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London. *Series A: Mathematical, Physical and Engineering Sciences, 454,* 903–995. https://doi.org/10.1098/rspa.1998.0193

Huang, Y. (2021). Hilbert-Huang transform for climate time series forecasting. *Journal of Atmospheric Research, 98*.

Ismail, A., & Oke, I. A. (2012). Trend analysis of precipitation in Birnin Kebbi, Nigeria. *International Research Journal of Agricultural Science and Soil Science, 2*(7), 286–297.

Jokubaitis, S., Celov, D., & Leipus, R. (2021). Sparse structures with LASSO through principal components: Forecasting GDP components in the short-run. *International Journal of Forecasting, 37*(2), 759–776. https://doi.org/10.1016/j.ijforecast.2020.09.005

Jungbacker, B., & Koopman, S. J. (2015). Likelihood-based dynamic factor analysis for measurement and forecasting. *The Econometrics Journal, 18*(1), C1–C21.

Kajtar, J. B., Santoso, A., Collins, M., Taschetto, A. S., England, M. H., & Frankcombe, L. M. (2021). CMIP5 intermodel relationships in the baseline Southern Ocean climate system and with future projections. *Earth's Future, 9*(6), 1–21. https://doi.org/10.1029/2020EF001873

Khesali, E., & Mobasheri, M. R. (2023). Near surface air temperature estimation through parametrization of MODIS products. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 10*. https://doi.org/10.5194/isprs-annals-X-4-W1-2022-405-2023

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *Proceedings of the 5th International Conference on Learning Representations.*

Knight, K., & Fu, W. (2000). Asymptotics for LASSO-type estimators. *The Annals of Statistics, 28*(5), 1356–1378.

Liu, Z. (2018). Vector autoregressive models for climate time series forecasting. *Journal of Atmospheric Science.*

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J., & West, M. (2006). Sparse statistical modelling in gene expression genomics. In K. A. Do, P. M., & M. Vannucci (Eds.), *Bayesian inference for gene expression and proteomics* (pp. 723–732). Cambridge University Press.

Magnano, B. (2008). Generation of stochastic weather sequences for climate impact assessments. *Journal of Climate, 25.*

Refsgaard, J. C., Arnbjerg-Nielsen, K., & Drews, M. (2013). The role of uncertainty in climate change adaptation strategies: A Danish water management example. *Mitigation and Adaptation Strategies for Global Change, 18,* 337–359. https://doi.org/10.1007/s11027-012-9366-6

Modarres, R., & Ouarda, T. (2013). Heteroscedasticity in time series modeling of climatic variables. *Journal of Hydrology.*

Modarres, R., & Ouarda, T. (2014). Modeling heteroscedasticity in precipitation data using GARCH models. *Journal of Hydrology.*

Mustapha, D. A., Adams, S. O., & Adehi, M. U. (2025). Dynamic time series models for forecasting climatic

variables. *Journal of Statistics, 15*(1), 1–13. https://doi.org/10.5923/j.statistics.20251501.01

Niclos, R., Valiente, J. A., Barbera, M. J., & Caselles, V. (2014). Land surface air temperature retrieval from EOS-MODIS images. *IEEE Geoscience and Remote Sensing Letters, 11*. https://doi.org/10.1109/LGRS.2013.2293540

Nieto, H., Sandholt, I., Aguado, I., Chuvieco, E., & Stisen, S. (2011). Air temperature estimation with MSG-SEVIRI data: Calibration and validation of the TVX algorithm for the Iberian Peninsula. *Remote Sensing of Environment, 115*. https://doi.org/10.1016/j.rse.2010.08.010

Niles, M. T., Brown, M., & Dynes, R. (2016). Farmer's intended and actual adoption of climate change mitigation and adaptation strategies. *Climatic Change, 135,* 277–295. https://doi.org/10.1007/s10584-015-1558-0

Ninyerola, M., Pons, X., & Roure, J. (2007). Monthly precipitation mapping of the Iberian Peninsula using spatial interpolation tools implemented in a geographic information system. *Theoretical and Applied Climatology, 89,* 195–209. https://doi.org/10.1007/s00704-006-0264-2

Okhakhu, P. A. (2010). *The significance of climatic elements in planning the urban environment of Benin City, Nigeria* (Unpublished doctoral dissertation). Ambrose Alli University.

Oyediran, O. (1977). *The climates of West Africa*. Heinemann Education Books.

Park, T. (2023). Real-time prediction for climate using multivariate time series. *Journal of Applied Earth Science.*

Prihodko, L., & Goward, S. N. (1997). Estimation of air temperature from remotely sensed surface observations. *Remote Sensing of Environment, 60*. https://doi.org/10.1016/S0034-4257(96)00216-7

Qaisrani, Z. N., Nuthammachot, N., & Techato, K. (2021). Drought monitoring based on standardized precipitation index and standardized precipitation evapotranspiration index in the arid zone of Balochistan province, Pakistan. *Arabian Journal of Geosciences*. https://doi.org/10.1007/s12517-020-06302-w

Ragatoa, D. S., Ogunjobi, K. O., Okhimamhe, A. A., Francis, S. D., & Adet, L. (2018). A trend analysis of temperature in selected stations in Nigeria using three different approaches. *Open Access Library Journal, 5*(2), 1–17. https://doi.org/10.4236/oalib.1104371

Slater, L. J., Arnal, L., & Boucher, M.-A. (2023). Hybrid forecasting: Blending climate predictions with AI models. *Hydrology and Earth System Sciences, 27,* 1865–1889.

Stock, J. H., & Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics, 20,* 147–162.

Van der Wiel, K., & Bintanja, R. (2021). Contribution of climatic changes in mean and variability to monthly temperature and precipitation extremes. *Communications Earth & Environment, 2*(1), 1–21. https://doi.org/10.1038/s43247-020-00077-4

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30.*

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Statistics, 7,* 723–732.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association, 101*(476), 1418–1429.