

ME5311 Project Report: Data-Driven Analysis of Indo-Pacific Climate Using PCA, DMD, and Deep Learning*

Linjie Wang^a

^aDepartment of Mechanical Engineering, National University of Singapore

This manuscript was compiled on April 17, 2025

1 This project presents a data-driven approach for analyzing and forecasting climate variability over the Indo-Pacific region using a combination
2 of Principal Component Analysis (PCA), Dynamic Mode Decomposition (DMD), and Long Short-Term Memory (LSTM) networks. We begin by
3 applying PCA to reduce the spatial dimensionality of high-resolution sea level pressure (SLP) and two-meter temperature (T2M) datasets,
4 retaining the dominant modes that capture 95% of the total variance. The reduced datasets are then used to construct linear dynamical
5 systems using DMD, enabling modal analysis and short-term single-step forecasting. To capture nonlinear temporal dependencies, LSTM
6 models are further employed for sequence-to-one time series prediction. Comparative evaluations using RMSE, MAE, and R² demonstrate that
7 LSTM consistently outperforms DMD in predictive accuracy, especially in leading PCA modes. These results underscore the advantages of
8 combining linear decomposition techniques with deep learning for interpretable and accurate climate forecasting.

Data-Driven Modeling | Climate Forecasting | Principal Component Analysis | Dynamic Mode Decomposition | LSTM

1 Introduction

2 Data-driven engineering and machine learning have been applied in various engineering domains, including structural, mechanical,
3 civil and aerospace engineering. In structural engineering, machine learning has been used to predict the behavior of structures
4 under different loads and environmental conditions. In mechanical engineering, machine learning has been used to optimize the
5 design of components and systems, and to predict their performance under different operating conditions. In civil engineering,
6 machine learning has been used to predict the behavior of soils and structures, and to optimize construction processes. In aerospace
7 engineering, machine learning has been used to predict the performance of aircraft and spacecraft, and to optimize
8 their design and operation.

9 Climate systems exhibit inherent complexity characterized by multi-scale interactions, high dimensionality, nonlinearity,
10 and chaotic dynamics, presenting significant challenges for accurate analysis and prediction (1). Traditional climate models
11 typically rely on numerically solving partial differential equations governing atmospheric and oceanic dynamics; however, these
12 approaches often encounter limitations due to high computational costs, constrained spatial resolutions, and inaccuracies
13 arising from parameterization errors. Furthermore, the nonlinear interactions and multiscale coupling intrinsic to climate
14 processes exacerbate these difficulties, limiting the reliability of long-term predictions. Recent advancements in data-driven
15 methodologies (2), particularly machine learning (ML) techniques (3), offer substantial advantages for climate modeling. These
16 techniques have demonstrated exceptional capabilities in processing vast climate datasets, capturing complex spatiotemporal
17 dependencies, and enhancing predictive performance, all while significantly reducing computational demands. Nonetheless,
18 the interpretability of ML-based models remains a critical issue, often reducing their practical credibility. To overcome this
19 limitation, interpretable machine learning frameworks are increasingly employed, improving transparency and facilitating
20 deeper insights into the underlying physical dynamics captured by these data-driven models (4).

21 Dimensionality reduction is a critical technique in data analysis and machine learning, particularly when dealing with
22 high-dimensional datasets such as those encountered in climate modeling. High-dimensional data can lead to several challenges,
23 including increased computational complexity, overfitting, and difficulties in data visualization and interpretation. By reducing
24 the number of input variables or features, dimensionality reduction helps mitigate these issues, leading to more efficient and
25 effective data processing (5). Several mainstream techniques are employed for dimensionality reduction. Principal Component
26 Analysis (PCA) (6) is a widely used linear method that transforms the original variables into a new set of uncorrelated
27 variables (principal components) ordered by the amount of variance they capture from the data. Non-linear techniques like
28 t-Distributed Stochastic Neighbor Embedding (t-SNE) (7) and Uniform Manifold Approximation and Projection (UMAP) (8)
29 are also popular, especially for visualizing complex data structures. t-SNE emphasizes the local structure and is effective in
revealing clusters, while UMAP preserves the local and global data structure and is computationally more efficient. Although

*The code for this project is available at <https://github.com/LINJIE-WANG/ME5311-Project-Data-Driven-Climate-Modeling.git>

dimensionality reduction is valuable, it poses challenges such as selecting the right method and number of dimensions, which can greatly influence results. While dimensionality reduction offers significant benefits, it also presents challenges, including the selection of appropriate techniques and the determination of optimal dimensionality, both of which can substantially affect outcomes. In addition, the risk of information loss can impact downstream analyses. In this study, methods such as PCA are used to simplify high-dimensional climate data, facilitating more efficient modeling and enhancing the accuracy of weather system forecasts.

Dynamic Mode Decomposition (DMD) (9) is a data-driven technique that decomposes complex dynamical systems into a set of modes, each associated with specific temporal behaviors, such as growth, decay, or oscillation. DMD captures both spatial and temporal dynamics, making it particularly effective for analyzing high-dimensional time-series data. In recent years, DMD has been increasingly applied in climate science to extract modes of variability from atmospheric and oceanic datasets. For instance, DMD with control (DMDc) (10) has been utilized to separate the effects of climate dynamics from external forcings, such as emissions, in surface air temperature data. Despite its strengths, DMD faces challenges, particularly in handling noisy data and capturing nonlinear dynamics. To address these issues, various extensions of DMD have been developed. For example, Multi-Resolution DMD (mrDMD) (11) allows for the analysis of data across multiple time scales, enhancing the method's ability to handle transient behaviors. Additionally, Total Least Squares DMD (12) improves robustness against noise by incorporating regularization techniques. In this study, DMD is utilized to model the dynamics of key climate variables. The extracted modes are analyzed to evaluate system stability and identify dominant patterns governing weather behavior. Additionally, DMD enables single-step short-term forecasting by projecting the current state onto these modes, offering insights into near-future system evolution and enhancing forecast accuracy.

Long Short-Term Memory (LSTM) networks (13), a specialized form of recurrent neural networks (RNNs), are adept at modeling sequential data by capturing long-term dependencies through their unique gating mechanisms. This capability makes them particularly suitable for weather forecasting, where temporal patterns and nonlinear relationships among variables like temperature, humidity, and precipitation are prevalent. Recent research has demonstrated the efficacy of LSTM models in various meteorological applications. For instance, LSTM networks have been employed to predict daily climate variables such as temperature, precipitation, and humidity, achieving satisfactory levels of accuracy in temperature forecasting (14). Additionally, hybrid models combining Convolutional Neural Networks (CNNs) and LSTMs have been developed to enhance the prediction of extreme weather events, such as floods, by effectively capturing both spatial and temporal features (15). Moreover, integrating decomposition methods like Seasonal Decomposition (SD) and Ensemble Empirical Mode Decomposition (EEMD) with CNN-LSTM architectures has shown improved accuracy in multi-step forecasts of meteorological variables, including temperature, wind speed, and precipitation (16). In our project, we utilize LSTM networks to perform time series forecasting of meteorological variables, leveraging their strength in modeling complex temporal dependencies to improve the accuracy of short-term weather predictions.

The remainder of this paper is organized as follows. The next section details the application of Principal Component Analysis (PCA) for reducing the spatial dimensionality of the climate dataset. This is followed by a section on Dynamic Mode Decomposition (DMD), which constructs a linear dynamical system using the PCA-reduced data to enable system analysis and single-step short-term forecasting. Subsequently, we present the use of Long Short-Term Memory (LSTM) networks to model nonlinear temporal dependencies and perform data-driven time-series prediction. Finally, the paper concludes with a summary of the main findings and directions for future work.

Dimensionality Reduction Using PCA

In this project, we use a large-scale climate reanalysis dataset covering the Indo-Pacific region, with a spatial resolution of $0.5^\circ \times 0.5^\circ$ and daily temporal resolution from December 31, 1979, to December 31, 2022, totaling 16,071 days. The dataset includes two key atmospheric variables: sea level pressure (SLP) and two-meter temperature (T2m). Each daily snapshot contains a spatial grid of 161×101 , resulting in 16,261 spatial dimensions. The high dimensionality of the data poses significant computational challenges and increases the risk of overfitting in dynamical analysis and predictive modeling. Therefore, we apply Principal Component Analysis (PCA) to reduce the spatial dimensionality of the dataset, extracting dominant spatial modes that capture the main variance in the data.

Mathematical Foundation of PCA. Principal Component Analysis (PCA) aims to perform a linear transformation that projects high-dimensional data into a lower-dimensional subspace while preserving as much variance as possible. Given a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, where n is the number of samples and p is the number of features, the standard procedure of PCA includes the following steps:

1. **Data Standardization:** Center the data by subtracting the mean of each feature and optionally scale it to unit variance to eliminate the influence of different feature magnitudes:

$$\tilde{\mathbf{X}} = \mathbf{X} - \bar{\mathbf{X}} \quad [1]$$

2. **Covariance Matrix Computation:** Calculate the covariance matrix of the standardized data:

$$\mathbf{C} = \frac{1}{n-1} \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \quad [2]$$

86 3. **Eigen Decomposition:** Perform eigenvalue decomposition on the covariance matrix to obtain eigenvalues and their
 87 corresponding eigenvectors:

$$\mathbf{C} = \mathbf{Q}\Lambda\mathbf{Q}^\top \quad [3]$$

89 where Λ is a diagonal matrix of eigenvalues and \mathbf{Q} contains the corresponding eigenvectors.

90 4. **Dimensionality Reduction:** Select the top k eigenvectors corresponding to the largest eigenvalues to form the
 91 transformation matrix $\mathbf{W} \in \mathbb{R}^{p \times k}$. Then project the original data into the reduced space:

$$\mathbf{Z} = \tilde{\mathbf{X}}\mathbf{W} \quad [4]$$

93 where \mathbf{Z} is the lower-dimensional representation of the data.

94 This process effectively reduces the dimensionality of the original data while preserving its most significant variance structure.

95 **PCA Implementation.** We first flatten the spatial fields of SLP and T2M into one-dimensional vectors and reconstruct the
 96 dataset into a matrix of size 16071×16261 , where each row represents a daily sample and each column corresponds to a spatial
 97 grid point. Following the standard PCA procedure, we perform dimensionality reduction based on the cumulative explained
 98 variance ratio. The threshold is set to 95%, and only the principal components accounting for up to 95% of the total variance
 99 are retained. The cumulative explained variance ratio is defined as:

$$\eta_k = \frac{\sum_{i=1}^k \lambda_i}{\sum_{j=1}^p \lambda_j} \quad [5]$$

100 where λ_i denotes the eigenvalue corresponding to the i -th principal component, p is the total number of components, and η_k
 101 represents the proportion of variance explained by the first k components.

102 To achieve the 95% cumulative explained variance threshold, 15 principal components are required for SLP and 10 for T2M.
 103 For consistency in subsequent analysis and model comparison, we uniformly retain 15 principal components for both variables,
 104 resulting in a reduced dataset of size 16071×15 . As presented in Table 1, the cumulative explained variance reaches 95.28% for
 105 SLP and 95.25% for T2M within the first 15 and 10 modes, respectively, supporting the selection of 15 components across both
 106 fields. Figure 1 visualizes the dominant spatial patterns captured by PCA, with panels (A–D) showing the first four principal
 107 modes of SLP and panels (E–H) showing those of T2M. The first PCA mode of SLP (panel A), which accounts for 66.90% of
 108 the total variance, reveals a large-scale pressure anomaly pattern over the Indo-Pacific region that may reflect the influence of
 109 tropical convection or monsoon-related variability. Similarly, the leading mode of T2M (panel E), explaining 84.69% of the
 110 total variance, captures a broad temperature gradient structure, which may be associated with large-scale land-sea thermal
 111 contrast and surface heating over the oceanic domain.

Table 1. Explained variance and cumulative variance of the first 15 PCA modes for SLP and T2M

| PCA Mode | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | |
|----------|----------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| SLP | Variance (%) | 66.90 | 6.85 | 4.63 | 4.36 | 2.46 | 2.30 | 2.16 | 1.36 | 1.11 | 0.82 | 0.62 | 0.52 | 0.47 | 0.40 | 0.32 |
| | Cumulative (%) | 66.90 | 73.75 | 78.38 | 82.74 | 85.20 | 87.50 | 89.66 | 91.02 | 92.13 | 92.95 | 93.57 | 94.09 | 94.56 | 94.96 | 95.28 |
| T2M | Variance (%) | 84.69 | 4.87 | 1.50 | 1.26 | 0.86 | 0.56 | 0.46 | 0.41 | 0.33 | 0.31 | 0.25 | 0.20 | 0.17 | 0.16 | 0.14 |
| | Cumulative (%) | 84.69 | 89.56 | 91.06 | 92.32 | 93.18 | 93.74 | 94.20 | 94.61 | 94.94 | 95.25 | 95.50 | 95.70 | 95.87 | 96.03 | 96.17 |

113 Dynamical System Modeling and Short-Term Prediction via DMD

114 In this section, we employ the PCA-reduced climate data to construct a linear dynamical system using Dynamic Mode
 115 Decomposition (DMD). As a data-driven technique, DMD extracts spatiotemporal coherent structures from high-dimensional
 116 time-series data by identifying dynamic modes and their associated eigenvalues. Unlike purely statistical methods, DMD
 117 captures both spatial patterns and temporal characteristics, including mode-specific frequencies and growth rates, enabling
 118 physically interpretable system analysis. In this study, we compute the full spectrum of DMD eigenvalues, extract dominant
 119 modes, and quantify their temporal behaviors. The resulting model is further used to perform short-term, single-step forecasting
 120 of climate variables.

121 **Mathematical Foundation of DMD.** Mathematically, DMD seeks a linear operator that approximates the evolution of a dynamical
 122 system from time-resolved data. The following describes the standard formulation of DMD. Consider a time series of snapshots
 123 $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$, where each $\mathbf{x}_i \in \mathbb{R}^n$ represents the system state at time t_i . We define two data matrices:

$$\mathbf{X}_1 = [\mathbf{x}_1 \ \mathbf{x}_2 \ \dots \ \mathbf{x}_{m-1}], \quad \mathbf{X}_2 = [\mathbf{x}_2 \ \mathbf{x}_3 \ \dots \ \mathbf{x}_m]. \quad [6]$$

124 We seek a linear operator \mathbf{A} such that:

$$\mathbf{X}_2 \approx \mathbf{A}\mathbf{X}_1. \quad [7]$$

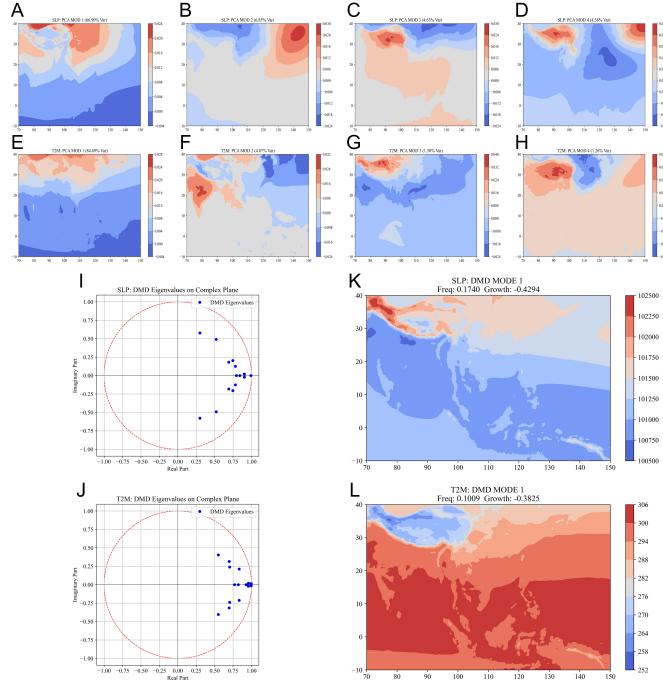


Fig. 1. Principal component analysis (PCA) and dynamic mode decomposition (DMD) results of sea level pressure (SLP) and 2-meter temperature (T2M). (A–D) The first four spatial modes of SLP from PCA. (E–H) The first four spatial modes of T2M from PCA. (I, J) DMD eigenvalues on the complex plane for SLP and T2M, respectively. (K, L) The first DMD modes of SLP and T2M, including spatial structures and corresponding frequency and growth rate.

Since $\mathbf{A} \in \mathbb{R}^{n \times n}$ is typically high-dimensional and not computed directly, we apply a low-rank approximation using the Singular Value Decomposition (SVD) of \mathbf{X}_1 :

$$\mathbf{X}_1 = \mathbf{U}\Sigma\mathbf{V}^T, \quad [8]$$

where $\mathbf{U} \in \mathbb{R}^{n \times r}$, $\Sigma \in \mathbb{R}^{r \times r}$, and $\mathbf{V} \in \mathbb{R}^{(m-1) \times r}$, with r being the truncated rank.

The projected low-dimensional linear operator is given by:

$$\tilde{\mathbf{A}} = \mathbf{U}^T \mathbf{X}_2 \mathbf{V} \Sigma^{-1}. \quad [9]$$

We then compute the eigen-decomposition of $\tilde{\mathbf{A}}$:

$$\tilde{\mathbf{A}}\mathbf{W} = \mathbf{W}\Lambda, \quad [10]$$

where Λ contains the eigenvalues and \mathbf{W} contains the eigenvectors. The DMD modes in the original high-dimensional space are reconstructed as:

$$\Phi = \mathbf{X}_2 \mathbf{V} \Sigma^{-1} \mathbf{W}. \quad [11]$$

Each eigenvalue λ_i in Λ relates to a continuous-time growth rate and oscillation frequency via:

$$\omega_i = \frac{\log(\lambda_i)}{\Delta t}, \quad [12]$$

where $\omega_i = \alpha_i + i\beta_i$, with α_i indicating growth/decay rate and β_i representing oscillation frequency.

DMD thus provides a set of modes Φ and their corresponding temporal behaviors λ_i , offering a powerful tool for analyzing linear approximations of nonlinear dynamical systems.

DMD-Based System Identification. In this subsection, we construct a linear dynamical system using Dynamic Mode Decomposition (DMD) based on the PCA-reduced climate data. Since the data dimensionality has been reduced to 15 through PCA, all 15 DMD modes are retained in the model to preserve the complete dynamic information in the reduced space.

The DMD algorithm yields a set of eigenvalues characterizing the temporal evolution of each mode. The magnitudes of all computed eigenvalues are found to be less than one, indicating that both the SLP and T2M systems are dynamically stable under the linear approximation. These eigenvalue spectra are visualized in Figure 1 panels (I) and (J) for SLP and T2M, respectively. Furthermore, the DMD modes corresponding to the leading eigenvalue (i.e., the eigenvalue with the largest magnitude) are presented in panels (K) and (L), illustrating the dominant spatial structures governing the short-term behavior

of each system. To further quantify the temporal behavior, Table 2 lists the frequencies and growth rates of the dominant DMD modes. The growth rates for all listed modes are negative, reinforcing the conclusion that the system dynamics are asymptotically stable. Moreover, the relatively low frequency values and energy concentration in the leading modes suggest that the system evolution is governed by a few coherent structures, supporting the potential for accurate short-term prediction within this reduced-order framework.

Table 2. Selected DMD mode frequencies and growth rates for SLP and T2M. For conjugate symmetric eigenvalue pairs, only one representative mode is shown.

| | DMD Mode | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|-----|-----------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| SLP | Frequency | 0.1740 | 0.1204 | 0.0409 | 0.0426 | 0.0000 | 0.0255 | 0.0000 | 0.0000 | 0.0037 |
| | Growth | -0.4294 | -0.3341 | -0.3363 | -0.2563 | -0.0087 | -0.2344 | -0.2287 | -0.1735 | -0.1013 |
| T2M | Frequency | 0.1009 | 0.0675 | 0.0522 | 0.0397 | 0.0000 | 0.0000 | 0.0024 | 0.0383 | 0.0000 |
| | Growth | -0.3825 | -0.2689 | -0.2964 | -0.1517 | -0.2608 | -0.1968 | -0.0051 | -0.0423 | -0.0762 |

Single-Step Forecasting Using DMD. To evaluate the predictive performance of the DMD-constructed dynamical system, we conduct short-term single-step forecasting, where the state at time t_{k+1} is predicted based on the state at time t_k using Equation 7. In this experiment, 80% of the time series data are used for model construction, while the remaining 20% are reserved for validation.

Table 3 summarizes the forecasting performance for the first five PCA modes, as well as the mean statistics across all DMD modes. For each mode, we compute three standard metrics: root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2). The results show that the leading PCA modes generally yield better predictive accuracy, with RMSE and MAE decreasing and R^2 increasing for the lower-indexed modes. This trend is particularly evident in the first mode of both SLP and T2M, where the R^2 exceeds 95%. The average R^2 across all modes reaches 63% for SLP and 71% for T2M, indicating that the DMD-based system provides reasonably good short-term predictive skill, especially in the dominant components.

Table 3. Comparison of single-step forecasting performance between DMD and LSTM models on the first five PCA modes and overall mean.

| | PCA Mode | 1 | 2 | 3 | 4 | 5 | Mean value | |
|------|----------|-----------|---------|---------|---------|---------|------------|---------|
| DMD | SLP | RMSE | 9781.55 | 7844.98 | 5831.92 | 6113.54 | 5145.90 | 4742.16 |
| | SLP | MAE | 7515.65 | 6171.96 | 4651.22 | 4690.56 | 3965.78 | 3700.29 |
| | SLP | R^2 (%) | 96.00 | 76.00 | 80.00 | 75.00 | 72.00 | 63.00 |
| | T2M | RMSE | 31.77 | 20.14 | 27.41 | 27.67 | 24.83 | 21.50 |
| | T2M | MAE | 24.65 | 15.71 | 21.30 | 21.30 | 19.62 | 16.77 |
| | T2M | R^2 (%) | 100.00 | 98.00 | 86.00 | 85.00 | 79.00 | 71.00 |
| LSTM | SLP | RMSE | 8916.75 | 5907.09 | 4485.92 | 4705.81 | 4204.40 | 3791.38 |
| | SLP | MAE | 6878.96 | 4557.07 | 3475.93 | 3619.77 | 3246.48 | 2943.24 |
| | SLP | R^2 (%) | 97.00 | 87.00 | 88.00 | 85.00 | 81.00 | 77.00 |
| | T2M | RMSE | 54.35 | 20.28 | 21.60 | 23.63 | 19.11 | 20.15 |
| | T2M | MAE | 42.79 | 15.87 | 16.61 | 17.77 | 15.70 | 15.71 |
| | T2M | R^2 (%) | 99.00 | 98.00 | 91.00 | 89.00 | 87.00 | 78.50 |

Nonlinear Sequence Forecasting Using LSTM

In this section, we employ Long Short-Term Memory (LSTM) networks to model and predict the nonlinear temporal dynamics of climate variables. While the previous section utilized DMD to construct a linear dynamical system for short-term forecasting, LSTM provides a more flexible framework capable of capturing nonlinear relationships and long-range dependencies in the data. The forecasting performance of the LSTM model is subsequently evaluated and compared with that of the DMD-based single-step prediction.

LSTM Model Architecture and Training Procedure. To train the LSTM forecasting model, the dataset was first divided into a training set (80%) and a testing set (20%). Prior to training, all input features were standardized to have zero mean and unit variance. The forecasting task was formulated as a supervised learning problem, where sequences of 10 consecutive states were used to predict the state at the next time step.

The model architecture consists of a single-layer Long Short-Term Memory (LSTM) network with 128 units, followed by a dropout layer and a fully connected output layer. This sequence-to-one architecture is designed for multivariate time-series forecasting. The model was trained using Adam optimizer, with mean squared error (MSE) as the loss function and a batch

size of 32. To mitigate overfitting, an early stopping strategy was employed. Specifically, training was halted if validation loss did not improve for eight consecutive epochs, ensuring efficient convergence and better generalization performance.

Forecasting Results and Performance Evaluation. To assess the predictive capability of the proposed models, we evaluate the single-step forecasting performance of both DMD and LSTM approaches using the first five principal components derived from PCA. Table 3 presents a comparative summary of the prediction accuracy in terms of RMSE, MAE, and R^2 for both SLP and T2M datasets. While the table reports results for the first five modes, the reported mean statistics are computed across all 15 PCA modes to provide a comprehensive evaluation.

Overall, the LSTM model outperforms DMD across all metrics and modes. For SLP, the average RMSE and MAE values decrease from 4742.16 and 3700.29 (DMD) to 3791.38 and 2943.24 (LSTM), respectively, while the mean R^2 improves from 63.00% to 77.00%. Similarly, for T2M, LSTM achieves lower error and higher R^2 , with the mean R^2 increasing from 71.00% (DMD) to 78.50% (LSTM). These results suggest that LSTM provides more accurate short-term forecasting, likely due to its ability to model nonlinear dependencies. Figure 2 visualizes the predicted versus true values of the first principal component for both SLP and T2M using DMD and LSTM models. Panels (A) and (B) correspond to the DMD predictions, while panels (C) and (D) show the results from the LSTM model. In both cases, LSTM demonstrates a better match with the ground truth, particularly in capturing amplitude and phase over time.

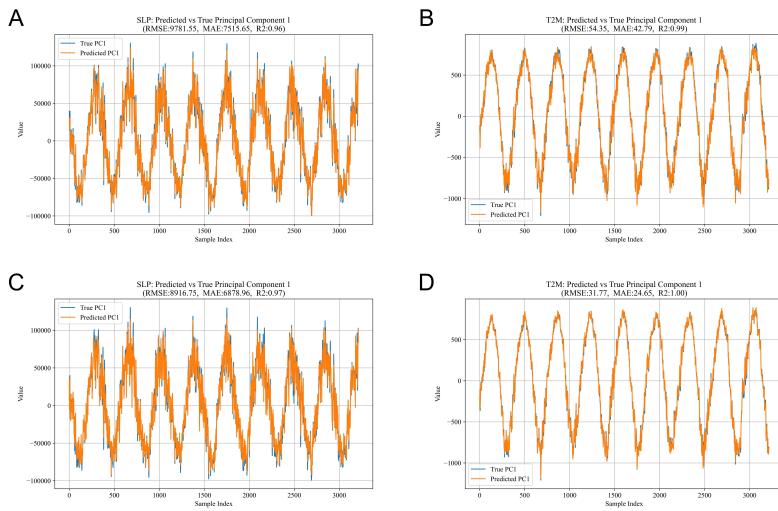


Fig. 2. Comparison of predicted versus true values of the first PCA mode for SLP and T2M using DMD (A–B) and LSTM (C–D) models.

Conclusion

In this study, we proposed an integrated data-driven framework that combines Principal Component Analysis (PCA), Dynamic Mode Decomposition (DMD), and Long Short-Term Memory (LSTM) networks to analyze and forecast key climate variables over the Indo-Pacific region. PCA was applied to reduce the spatial dimensionality of high-resolution climate data while preserving dominant spatial patterns. Based on the reduced-order representation, DMD enabled the construction of interpretable linear dynamical systems and provided insights into modal behaviors and system stability. However, the forecasting performance of DMD was limited due to the intrinsic linearity of the method. In contrast, LSTM demonstrated superior short-term prediction accuracy by effectively capturing nonlinear temporal dependencies in the data. Comparative evaluations using multiple performance metrics confirmed the advantages of LSTM over DMD for both sea level pressure (SLP) and two-meter temperature (T2M) predictions. Future work will explore hybrid modeling strategies that integrate the interpretability and modal insight of DMD with the flexibility and learning capacity of deep learning models, aiming to further improve forecast accuracy and physical realism in climate modeling.

ACKNOWLEDGMENTS. I would like to express my sincere gratitude to Prof. Gianmarco Mengaldo for his exceptional teaching and guidance throughout this course. Under his instruction, I not only gained practical skills in data processing but also developed a foundational understanding of dynamical systems and the power of machine learning techniques. What impressed me most about Prof. Mengaldo's teaching style is his problem-driven and application-oriented approach. Rather than focusing solely on abstract mathematical formulations, he always connects theory with real-world contexts, making the learning experience both engaging and enlightening. He frequently provides well-structured code examples and demonstrates real-world implementations in class, which greatly enhanced our understanding of the concepts. The opportunity to attend a guest lecture by a senior expert from NVIDIA further broadened my perspective and introduced me to cutting-edge developments in the industry. This course has significantly sharpened my data sensitivity and provided a more structured approach to modeling and analysis. I am truly grateful to Prof. Mengaldo for his dedication and inspiration, as well as to the TAs who supported the course throughout. The code and data used in this study are available at: <https://github.com/LINJIE-WANG/ME5311-Project-Data-Driven-Climate-Modeling.git>.

218 **References.**

- 219 1. S Bordini, S Kang, TA Shaw, I Simpson, L Zanna, The futures of climate modeling. *npj Clim. Atmospheric Sci.* **8**, 99 (2025).
220 2. Y Wu, W Xue, Data-driven weather forecasting and climate modeling from the perspective of development. *Atmosphere* **15**, 689 (2024).
221 3. CO de Burgh-Day, T Leeuwenburg, Machine learning for numerical weather and climate modelling: a review. *Geosci. Model. Dev.* **16**, 6433–6477 (2023).
222 4. R Yang, et al., Interpretable machine learning for weather and climate prediction: A review. *Atmospheric Environ.* p. 120797 (2024).
223 5. N Oskolkov, Dimensionality reduction: Overview, technical details, and some applications. *Appl. Data Sci. Tour. Interdiscip. Approaches, Methodol. Appl.* pp. 151–167 (2022).
224 6. H Abdi, LJ Williams, Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* **2**, 433–459 (2010).
225 7. L Van der Maaten, G Hinton, Visualizing data using t-sne. *J. machine learning research* **9** (2008).
226 8. L McInnes, J Healy, J Melville, Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
227 9. PJ Schmid, Dynamic mode decomposition of numerical and experimental data. *J. fluid mechanics* **656**, 5–28 (2010).
228 10. JL Proctor, SL Brunton, JN Kutz, Dynamic mode decomposition with control. *SIAM J. on Appl. Dyn. Syst.* **15**, 142–161 (2016).
229 11. JN Kutz, X Fu, SL Brunton, NB Erichson, Multi-resolution dynamic mode decomposition for foreground/background separation and object tracking in 2015 IEEE international conference on computer
230 vision workshop (ICCVW). (IEEE), pp. 921–929 (2015).
231 12. MS Hemati, CW Rowley, EA Deem, LN Cattafesta, De-biasing the dynamic mode decomposition for applied koopman spectral analysis of noisy datasets. *Theor. Comput. Fluid Dyn.* **31**, 349–368
232 (2017).
233 13. A Graves, A Graves, Long short-term memory. *Supervised sequence labelling with recurrent neural networks* pp. 37–45 (2012).
234 14. J Xu, Z Wang, X Li, Z Li, Z Li, Prediction of daily climate using long short-term memory (lstm) model. *Int. J. Innov. Sci. Res. Technol.* pp. 83–90 (2024).
235 15. V Atashi, HT Gorji, Enhanced flood prediction using lstm and climate parameters: multi-station analysis of snowmelt-induced flooding in the red river of the north. *J. Hydroinformatics* **27**, 245–260
236 (2025).
237 16. Q Guo, et al., A performance comparison study on climate prediction in weifang city using different deep learning models. *Water* **16**, 2870 (2024).

DRAFT