Transformer-Based Multi-Horizon Forecasting of the Niño 3.4 Index for ENSO Prediction

Jose Sulla-Torres[1]; Juan Diego Treviño-Ticlavilca[2]; Mario Cesar Santillana-Valdivia[3]; Andrea Sofía Maxi-Salas[4]; Robert Joaquin Medina-Ramos[5]
Universidad Católica de Santa María, Arequipa, Perú, [1,3,5]{jsullato, msantillana, robert.medina}@ucsm.edu.pe, [2,4]{juan.trevino, andrea.maxi}@estudiante.ucsm.edu.pe

Abstract*– El Niño–Southern Oscillation (ENSO) is a major source of interannual climate variability in the Pacific and affects water availability, agriculture, and disaster risk. We assessed whether a Transformer encoder driven only by four monthly oceanic indices can forecast the Niño 3.4 sea surface temperature anomaly one to three months ahead. Each input sequence contained 12 months of Niño 3.4, Niño 3, Niño 1+2, and the Oceanic Niño Index (ONI). We divided the series chronologically to preserve the forecasting setting: observations through 2017 were used for training, 2018–2021 for validation, and data from 2022 onward for testing. Across ten training seeds, the mean test RMSE was 0.351 ± 0.050 °C, which is used as the representative estimate of expected performance. The main deterministic run achieved an RMSE of 0.246 °C, an MAE of 0.202 °C, an R^2 of 0.907, and a Pearson correlation of 0.954; it had the lowest test RMSE among the ten seeds. The overall Critical Success Index (CSI) for detecting non-neutral ENSO months was 0.597. The phase-specific CSI was 0.892 for El Niño and 0.160 for La Niña; the latter result was associated with weak La Niña cases close to the classification threshold. Performance deteriorated mainly on the three-month horizon, although the Pearson correlation remained 0.933. The Diebold-Mariano test and block bootstrap showed statistically significant improvements over the persistence model in continuous metrics for the two- and three-month horizons, but not for the one-month horizon. McNemar's test showed the same significance pattern for categorical classification. Under the same protocol, we also trained an LSTM, a CNN-LSTM, and a Temporal Convolutional Network (TCN). The Transformer did not uniformly outperform all of them, although it achieved better aggregate continuous performance than persistence. These results indicate that a compact index-based model can retain useful short-range predictive skill; however, its tendency to underestimate anomalies across seeds, deterministic output, and limited predictor set restrict its direct operational use.

Pronóstico Multihorizonte del Índice Niño 3.4 mediante un Transformer Encoder para la Predicción de ENSO

Jose Sulla-Torres[1]; Juan Diego Treviño-Ticlavilca[2]; Mario Cesar Santillana-Valdivia[3]; Andrea Sofía Maxi-Salas[4]; Robert Joaquin Medina-Ramos[5]
Universidad Católica de Santa María, Arequipa, Perú, [1,3,5]{jsullato, msantillana, robert.medina}@ucsm.edu.pe, [2,4]{juan.trevino, andrea.maxi}@estudiante.ucsm.edu.pe

Resumen*– El Niño-Oscilación del Sur (ENSO) es una de las principales fuentes de variabilidad climática interanual en el Pacífico y afecta la disponibilidad hídrica, la agricultura y la gestión del riesgo. En este estudio se evaluó si un Transformer Encoder alimentado únicamente con cuatro índices oceánicos mensuales puede pronosticar la anomalía de la temperatura superficial del mar de Niño 3.4 con uno, dos y tres meses de anticipación. Cada entrada reunió 12 meses de Niño 3.4, Niño 3, Niño 1+2 y del Oceanic Niño Index (ONI). La serie se dividió cronológicamente para conservar un escenario de pronóstico real: entrenamiento hasta 2017, validación entre 2018 y 2021 y prueba desde 2022. En diez ejecuciones con diferentes semillas de entrenamiento, el RMSE promedio de prueba fue de 0.351 ± 0.050 °C, lo que se considera una estimación representativa del desempeño esperado. La corrida determinista principal obtuvo un RMSE de 0.246 °C, un MAE de 0.202 °C, un R^2 de 0.907 y una correlación de Pearson de 0.954; esta ejecución presentó el menor RMSE entre las diez semillas evaluadas. El Índice de Éxito Crítico (CSI) global para detectar meses de ENSO no neutrales fue de 0.597. Por fase, el CSI fue de 0.892 para El Niño y de 0.160 para La Niña; este último valor estuvo asociado a casos débiles de La Niña cercanos al umbral de clasificación. El desempeño se deterioró principalmente en el horizonte de tres meses, aunque la correlación de Pearson se mantuvo en 0.933. La prueba de Diebold-Mariano y el bootstrap por bloques mostraron mejoras estadísticamente significativas frente al modelo de persistencia en las métricas continuas para los horizontes de dos y tres meses, pero no para el de un mes. La prueba de McNemar mostró el mismo patrón de significancia en la clasificación categórica. Bajo el mismo protocolo se entrenaron también una LSTM, una CNN-LSTM y una red convolucional temporal (TCN). El Transformer no superó de manera uniforme a todas estas arquitecturas, aunque obtuvo un mejor desempeño agregado continuo que el de la persistencia. Los resultados indican que un modelo compacto basado en índices conserva una capacidad predictiva útil a corto plazo; sin embargo, su tendencia a subestimar las anomalías entre semillas, su naturaleza determinista y el reducido número de predictores limitan su aplicación operativa directa.

Palabras clave- ENSO, Niño 3.4, Transformer Encoder, aprendizaje profundo, temperatura superficial del mar, pronóstico multivariante.

I. Introducción

El Niño-Oscilación del Sur (ENSO) es la principal fuente de la variabilidad climática interanual en el Pacífico tropical. Su fase cálida, El Niño, y su fase fría, La Niña, emergen de la interacción no lineal entre el océano y la atmósfera y alteran la circulación, la precipitación y la temperatura en regiones muy alejadas de sus zonas de origen [1], [2]. El índice Niño 3.4, calculado sobre el Pacífico ecuatorial central, resume una parte sustancial de esta evolución y se utiliza ampliamente para el monitoreo y la verificación de pronósticos estacionales [3].

La anticipación de ENSO tiene valor científico y operativo. Los cambios de fase condicionan las decisiones sobre agricultura, recursos hídricos, energía y preparación ante emergencias; sin embargo, la predictibilidad varía entre décadas y estaciones, y suele deteriorarse cuando el horizonte temporal abarca la primavera boreal [3], [4]. En el Perú, además, los episodios cálidos intensos se relacionan con lluvias extremas, afectación de la infraestructura y movimientos en masa, por lo que una señal climática oportuna puede alimentar sistemas posteriores de pronóstico de precipitación y de susceptibilidad territorial [5], [6].

Los modelos dinámicos acoplados siguen siendo la base del pronóstico operacional de ENSO. Sin embargo, su costo de ejecución, la sensibilidad a las condiciones iniciales y los errores de parametrización han motivado el uso de métodos estadísticos y de aprendizaje profundo como herramientas complementarias. Estos métodos permiten explorar relaciones no lineales y obtener estimaciones rápidas a partir de series históricas. En particular, la atención resulta adecuada para relacionar meses no consecutivos dentro de una ventana multivariante sin imponer una estructura recurrente.

Este trabajo evalúa cómo un Transformer Encoder, alimentado con cuatro índices oceánicos mensuales, conserva la capacidad de pronosticar el índice Niño 3.4 con uno, dos y tres meses de anticipación. El experimento combina una salida multihorizonte directa, una partición cronológica sin mezcla aleatoria, métricas continuas y categóricas, y una comparación con persistencia. Con ello se busca estimar el aporte real de la arquitectura en una configuración compacta, sin atribuirle ventajas que no hayan sido contrastadas bajo el mismo protocolo.

A. Trabajos Relacionados

1) Pronóstico estadístico y aprendizaje profundo para ENSO: La predicción de ENSO se ha abordado mediante modelos dinámicos acoplados, métodos estadísticos y esquemas híbridos. El aprendizaje profundo amplió la capacidad de representar relaciones no lineales. Ham et al. emplearon redes neuronales convolucionales y aprendizaje de transferencia para obtener pronósticos multianuales de ENSO

[7]. Más recientemente, Wang et al. propusieron una arquitectura basada en Vision Transformer que integra información espaciotemporal en múltiples escalas [8]. Yan et al. combinaron la descomposición empírica y las redes neuronales temporales para predecir el Niño 3.4 y el Southern Oscillation Index (SOI) [9], mientras que Petersik y Dijkstra formularon pronósticos probabilísticos mediante redes neuronales [10]. Ham et al. incorporaron explícitamente la estacionalidad [11]. Hu et al. combinaron una red convolucional residual, *dropout* y aprendizaje de transferencia [12], mientras que Song et al. integraron pronósticos procedentes de modelos acoplados en una arquitectura residual profunda [13].

2) Atención y modelos multivariantes de nueva generación: La atención se ha incorporado progresivamente a los modelos climáticos. ENSO-GTC aprende teleconexiones espaciotemporales a escala global [14]. El modelo 3D-Geoformer utiliza autoatención para pronosticar campos tridimensionales de la temperatura superficial del mar (SST) y del esfuerzo del viento [15], y posteriormente se empleó para estudiar la prolongación de La Niña en 2021 [16]. En paralelo, los enfoques interpretables han buscado asociar la predicción con patrones térmicos físicamente plausibles [17]. ENSO-PhyNet integra procesos de balance de calor en la atención [18], y el uso conjunto de la SST y la salinidad superficial ha demostrado ser útil para ampliar el horizonte de predicción y mitigar parcialmente la barrera de la predictibilidad en primavera [19].

Los modelos citados emplean campos espaciales, información subsuperficial o restricciones físicas y, por ello, responden a objetivos distintos del planteado aquí. El presente estudio utiliza cuatro índices escalares y un horizonte temporal máximo de tres meses. La comparación se centra en determinar cuánto puede aportar la autoatención cuando la entrada ya resume la dinámica oceánica, sin pretender reemplazar los modelos climáticos de alta dimensión.

3) Transformers para series temporales: El Transformer original introdujo el mecanismo de atención de múltiples cabezas y el procesamiento paralelo de secuencias [20]. Para el pronóstico temporal, el Temporal Fusion Transformer (TFT) combinó atención e interpretación multihorizonte [21]; Informer redujo el costo de la atención en secuencias largas [22]; Autoformer incorporó descomposición y autocorrelación [23]; FEDformer trasladó parte del aprendizaje al dominio de la frecuencia [24]. Comparaciones recientes en pronóstico climático también han evaluado conjuntamente distintas arquitecturas Transformer [25]; y PatchTST mostró que la segmentación en parches puede mejorar la representación de series multivariadas [26]. No obstante, una secuencia de 12 meses no requiere mecanismos diseñados para miles de pasos. Por ello, se mantiene un codificador estándar (Encoder) y se centra la evaluación en la validez temporal, el sesgo y la degradación en función del horizonte.

TABLA I
SÍNTESIS DE ENFOQUES RECIENTES PARA EL PRONÓSTICO DE ENSO

Trabajo	Síntesis
Ham et al. [7]	CNN con campos de SST, de calor oceánico y de viento; transferencia de aprendizaje y pronóstico de largo plazo.
Yan et al. [9]	EEMD-TCN sobre Niño 3.4 y SOI; descomposición de series escalares para predicción multihorizonte.
Zhou y Zhang [15]	3D-Geoformer con temperatura oceánica y esfuerzo del viento; pronóstico multivariante.
Mu et al. [18]	Transformer guiado por el equilibrio térmico, la consistencia física y la interpretación del proceso.
Wang et al. [19]	SST y salinidad superficial; mejora de la habilidad en horizontes extendidos.
Este trabajo	Cuatro índices escalares: pronóstico directo a 3 meses con un conjunto reducido de predictores y evaluación basada en eventos.

La Tabla I permite distinguir entre dos escalas de modelación. Los trabajos orientados a horizontes largos incorporan campos espaciales y variables físicas adicionales; este estudio utiliza únicamente índices mensuales. Esa reducción facilita la integración del predictor en una cadena de procesamiento, pero también limita la información disponible para representar extremos y teleconexiones. Por esta razón, las correlaciones no se comparan directamente entre estudios, y el análisis se centra en la diferencia respecto de la persistencia dentro del mismo conjunto de datos.

El resto del artículo se organiza de la siguiente manera. La Sección II aborda la metodología, incluyendo los datos, la arquitectura y el protocolo experimental; la Sección III expone los resultados; la Sección IV presenta la discusión; y la Sección V presenta las conclusiones de la investigación.

II. METODOLOGÍA

A. *Fuente de datos y regiones oceánicas*

El experimento se construyó a partir del producto NOAA/NCEI Optimum Interpolation Sea Surface Temperature (OISST), versión 2.1, con una resolución espacial de 0.25° [27]. Para cada mes se calculó el promedio espacial de las regiones de la Tabla II.

TABLA II
REGIONES OCEÁNICAS UTILIZADAS COMO PREDICTORES

Índice	Definición
Niño 3.4	5°S-5°N, 190°-240°E (variable objetivo)
Niño 3	5°S-5°N, 210°-270°E
Niño 1+2	10°S-0°, 270°-280°E
ONI	Ídem Niño 3.4, suavizado (media móvil de 3 meses)

Niño 3.4 se definió como variable objetivo porque resume la variabilidad térmica del Pacífico ecuatorial central y constituye una referencia habitual para caracterizar el ENSO [1]. Niño 3 y Niño 1+2 aportan información de sectores progresivamente más orientales, mientras que el Oceanic Niño Index (ONI) representa una versión suavizada de la señal central. Como el ONI se deriva de Niño 3.4, ambos predictores no son independientes; se mantuvieron para comprobar si el suavizado de tres meses aportaba información útil. La salida objetivo, en cambio, se mantuvo sin suavizar para evaluar la anomalía mensual.

Las anomalías se expresaron en grados Celsius con respecto de la climatología de 1991-2020. Antes de formar las ventanas, se ordenaron los registros, se verificó la continuidad mensual y se eliminaron las observaciones incompletas. No se interpolaron datos en los periodos de validación y prueba para evitar que la información posterior influya en la evaluación.

B. Construcción de muestras y normalización

1) Suavizado: Se aplicó una media móvil causal de tres meses a las series oceánicas no suavizadas; el ONI, ya suavizado, no recibió un segundo filtrado.

2) Formulación supervisada: para cada instante t se construye una matriz de entrada con los 12 meses anteriores de los cuatro índices y un vector objetivo con Niño 3.4 en t, t+1 y t+2. El modelo produce los tres horizontes en una sola inferencia, evitando que el error del primer mes se propague recursivamente a los siguientes.

$$X_t = \left[x_{(t-12)}, \ldots, x_{(t-1)}\right] \in R^{(12\times 4)}, \quad y_t = \left[n_t, n_{(t+1)}, n_{(t+2)}\right] \in R^3 \qquad (1)$$

3) Normalización: cada predictor se estandariza utilizando la media y la desviación estándar estimadas únicamente a partir del conjunto de entrenamiento. Los mismos parámetros se aplican tanto a la validación como a la prueba. Esta decisión evita que la escala de los periodos futuros influya en el ajuste.

4) Partición temporal: el entrenamiento comprende las muestras cuyo objetivo finaliza el 31/12/2017; la validación cubre 2018-2021; y la prueba abarca desde 2022 hasta el último mes disponible. No se realiza mezcla aleatoria. La validación se utiliza para seleccionar el checkpoint y detener el aprendizaje, mientras que la prueba se mantiene aislada hasta la evaluación final.

La partición reproduce la situación de uso prevista: el modelo se ajusta a observaciones pasadas y se evalúa en un bloque futuro continuo. También evita que ventanas cercanas de un mismo episodio de ENSO aparezcan simultáneamente en los conjuntos de entrenamiento y de prueba, lo que podría inflar las métricas.

C. Arquitectura del Modelo

Se implementó un Transformer Encoder para regresión multivariante con entradas de forma (batch, 12, 4). Cada token corresponde a un mes y reúne los cuatro índices oceánicos. A partir de esa secuencia, el modelo genera tres valores de Niño 3.4, uno para cada horizonte de pronóstico.

1) Proyección y codificación posicional

Una capa densa proyecta cada vector de cuatro características en un espacio latente de dimensión d_model = 512. A esta proyección se suma una codificación posicional sinusoidal fija, necesaria porque la autoatención, por sí sola, no distingue el orden de los tokens [20]. La representación inicial puede expresarse como:

$$Z_0 = XW_e + b_e + P \qquad (2)$$

Donde W_e y b_e son los parámetros de la proyección y P contiene las posiciones de los 12 meses. Aunque el espacio latente es amplio en relación con el número de variables, se emplean dropout, normalización y detención temprana para mitigar el sobreajuste.

2) Bloques de autoatención

Se apilan seis bloques Encoder con el esquema Pre-LayerNorm [28]. Cada bloque contiene una atención multicabeza con 16 cabezas, 32 dimensiones por cabeza y una red feed-forward de 1024 unidades con activación GELU (Gaussian Error Linear Unit). Para una cabeza, la atención se calcula mediante:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (3)$$

Las 16 salidas se concatenan y se proyectan nuevamente al espacio de 512 dimensiones. Las conexiones residuales preservan la información de entrada, mientras que el dropout = 0.2 se aplica después de la atención y de la red feed-forward. Una normalización final estabiliza la representación acumulada.

3) Agregación temporal y salida multihorizonte

El promedio global de la dimensión temporal resume los 12 estados mensuales en un vector de 512 dimensiones. La cabeza de regresión está compuesta por Dense(256), GELU, dropout y Dense(3). Cada salida corresponde a un horizonte fijo, de modo que los tres meses se aprenden conjuntamente sin recurrir a predicciones recursivas.

4) Función de pérdida

La función objetivo fue el *Mean Squared Error* (MSE) calculado sobre las muestras y los tres horizontes. Esta pérdida impone una penalización mayor a los errores de gran magnitud, que son frecuentes durante las anomalías intensas. Los horizontes recibieron el mismo peso, por lo que el ajuste buscó un equilibrio entre el desempeño del primer, segundo y tercer mes.

$$L_{MSE} = (1/3N)\sum_i \sum_h \left(\hat{y}_{(i,h)} - y_{(i,h)}\right)^2 \qquad (4)$$

D. Entrenamiento y Protocolo de Evaluación

1) Configuración de entrenamiento: El ajuste se realizó con AdamW, que desacopla la regularización por decaimiento de pesos del gradiente de la pérdida [29]. La tasa inicial fue de 5×10^{-4}, el weight decay fue de 1×10^{-3} y el scheduler del coseno redujo gradualmente la tasa hasta 1×10^{-6}. El clipping de la norma a 0.5 limitó las actualizaciones inestables. Los hiperparámetros principales se presentan en la Tabla III.

TABLA III
HIPERPARÁMETROS DE ENTRENAMIENTO

Parámetro	Valor
Épocas máximas	200
Tamaño de batch	128
Gradient clipping (norma máx.)	0.5
Early stopping	paciencia 25 épocas, monitor val RMSE
Métrica de checkpoint	val RMSE
Scheduler	cosine annealing (LR min. 1x10^-6)
Semilla	42
Operaciones deterministas	habilitadas

La implementación empleó tf.data para la lectura, la creación de lotes y el prefetch. El checkpoint se seleccionó según el RMSE de validación y el entrenamiento se detuvo después de 25 épocas sin mejora. La semilla 42 definió la inicialización y el orden de los lotes de la corrida principal, como se muestra en las Tablas IV a VI. Para caracterizar la variabilidad del proceso de entrenamiento, el experimento se repitió íntegramente con diez semillas (42, 123, 456, 789, 1011, 1313, 1515, 1717, 2020 y 2323), manteniendo el resto del protocolo sin cambios; estos resultados se presentan en la Sección III.K. Todas las corridas se ejecutaron con semillas fijas en Python, NumPy y TensorFlow y con las operaciones deterministas de TensorFlow/cuDNN habilitadas (tf.config.experimental.enable_op_determinism), en la misma configuración de hardware y software (GPU NVIDIA RTX 5090, CUDA 12.5, cuDNN 9, TensorFlow 2.21). Al repetir la corrida principal se obtuvieron predicciones idénticas; por ello, la variabilidad entre semillas refleja únicamente la inicialización de los pesos y el orden de los lotes. Dado que el determinismo se garantiza solo dentro de una misma configuración, la ejecución en otro hardware (por ejemplo, en una CPU) puede producir resultados numéricamente distintos.

2) Métricas, clasificación y línea base: La evaluación continua incluyó MSE, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R^2, correlación de Pearson, sesgo medio y error máximo. RMSE y MAE cuantifican la magnitud del error en grados Celsius; R^2 mide la proporción de la variabilidad explicada; Pearson mide la asociación lineal entre la predicción y la observación; y el sesgo identifica una sobreestimación o subestimación persistente. Las métricas se analizaron en conjunto, ya que una alta correlación puede coexistir con errores de amplitud.

Para el análisis categórico se asignó El Niño cuando la anomalía fue > +0.5 °C, La Niña cuando fue < −0.5 °C y Neutral en el intervalo intermedio, de modo que el índice Niño 3.4 caracteriza ambas fases de ENSO. Precisión, Probability of Detection (POD), F1, False Alarm Ratio (FAR) y el Critical Success Index (CSI) se calcularon de forma binaria, considerando como evento cualquier mes no neutral y agregando los tres horizontes; el CSI se reportó además por separado para El Niño y La Niña. En paralelo, se reportó la exactitud de las tres fases por horizonte. Así, se separa la detección de un evento de la correcta identificación de su signo.

La línea base de persistencia supone que la última anomalía observada se mantiene constante durante los tres meses, $\hat{y}(t+h) = n(t-1)$. Es una referencia exigente en horizontes cortos debido a la memoria térmica del océano. La comparación de persistencia se evaluó mediante la correlación por horizonte, un test de Diebold-Mariano y un bootstrap por bloques sobre las métricas continuas (Sección III.J), y un test de McNemar sobre la clasificación categórica (Sección III.O). Adicionalmente, bajo el mismo protocolo, se contrastó el Transformer con tres arquitecturas de referencia — LSTM, CNN-LSTM y TCN — para situar su desempeño frente a otros modelos de aprendizaje profundo (Sección III.N), sin pretender establecer una comparación exhaustiva con variantes modernas de Transformer como TFT, Informer o PatchTST, que no fueron evaluadas.

III. Resultados

El modelo se evaluó sin reajuste durante la validación y la prueba. Se presentan resultados agregados, desempeño por horizonte, clasificación de fases y diagnósticos gráficos. Las métricas se calculan después de invertir la estandarización, por lo que los errores se expresan en grados Celsius.

A. Desempeño global

En la prueba, el modelo obtuvo un RMSE de 0.2456 °C y un MAE de 0.2016 °C. Los valores $R^2 = 0.9072$ y $r = 0.9539$ indican que se capturó una buena parte de la variabilidad temporal, aunque no se reprodujo con la misma precisión la amplitud de todas las anomalías. Tanto el error como la correlación fueron mejores en la prueba que en la validación, lo cual es coherente con la mayor variabilidad del periodo de prueba, dominado por El Niño 2023-2024, que favorece valores más altos de R^2 y de correlación. El sesgo global de −0.0235 °C es pequeño en esta corrida, pero varía con el horizonte (+0.063, −0.037 y −0.096 °C en los leads 1 a 3), y entre las diez semillas el sesgo medio de prueba fue de -0.105 ± 0.086 °C (Tabla VIII), negativo en nueve de ellas; esta tendencia a subestimar resulta especialmente relevante antes de considerar su uso operativo.

TABLA IV
Métricas Globales de Desempeño

Métrica	Validación	Prueba
MSE	0.0782	0.0603
RMSE (°C)	0.2796	0.2456
MAE (°C)	0.2114	0.2016
R^2	0.7872	0.9072
Correlación de Pearson	0.8929	0.9539
Sesgo (Bias, °C)	0.0366	−0.0235
Error máximo (°C)	0.8534	0.5807
Accuracy (no neutral)	0.8824	0.7807
Precisión	0.8947	0.8222
Recall (POD)	0.8095	0.6852
F1	0.8500	0.7475
FAR	0.1053	0.1778
CSI (no neutral)	0.7391	0.5968
CSI El Niño	0.8095	0.8919
CSI La Niña	0.6800	0.1600

B. Desempeño por horizonte

El error aumenta con el horizonte de validación: el RMSE pasó de 0.170 °C a un mes a 0.363 °C a tres meses. En prueba, el RMSE fue de 0.228, 0.195 y 0.301 °C para los leads 1, 2 y 3, de modo que el mayor error se concentró en el tercer mes. Las correlaciones de prueba se mantuvieron en 0.968, 0.971 y 0.933, y el MAE alcanzó 0.257 °C en el tercer mes. En otras palabras, el modelo conservó mejor la dirección general de la señal que su amplitud.

TABLA V
Desempeño por Mes de Predicción (Lead Time)

Conjunto	Lead	RMSE (°C)	MAE (°C)	Corr.
Val	1	0.170	0.132	0.964
Val	2	0.273	0.214	0.904
Val	3	0.363	0.288	0.807
Test	1	0.228	0.181	0.968
Test	2	0.195	0.167	0.971
Test	3	0.301	0.257	0.933

C. Clasificación de episodios ENSO

Con umbrales de ±0.5 °C, el CSI de meses no neutrales fue de 0.7391 en validación y de 0.5968 en prueba. En la prueba, la precisión (0.822) superó a la probabilidad de detección (0.685): la clasificación binaria produjo relativamente pocas falsas alarmas, pero omitió cerca de un tercio de los meses no neutrales. El desglose por fase muestra que este valor combina un desempeño alto para El Niño (CSI de 0.892) y uno bajo para La Niña (CSI de 0.160). Los casos de La Niña del bloque de prueba fueron débiles (anomalía media de −0.60 °C, próxima al umbral) y el modelo atenuó su amplitud (predicción media de −0.42 °C), por lo que 14 de los 18 casos se clasificaron como neutrales. En la validación, con episodios de La Niña más intensos (anomalía media de −0.90 °C), el CSI de esta fase fue de 0.680. La exactitud de las tres fases, por su parte, descendió de 84.2% en el primer mes al 68.4% en el tercero. Esta diferencia muestra que un pequeño error cerca del umbral puede alterar la fase asignada, aun cuando la trayectoria continua sea razonable.

TABLA VI
EXACTITUD DE CLASIFICACIÓN ENSO POR HORIZONTE

Conjunto	Lead 1	Lead 2	Lead 3
Val	91.2%	91.2%	82.4%
Test	84.2%	81.6%	68.4%

D. Diagrama de Dispersión

La Fig. 1 compara las predicciones con las observaciones para los tres horizontes. La mayoría de los puntos sigue la diagonal, aunque la dispersión aumenta en las anomalías de mayor magnitud. En el conjunto de prueba, las anomalías positivas más intensas (superiores a 1.5 °C) se subestiman levemente (error medio de −0.08 °C), y las anomalías negativas tienden a predecirse con menor intensidad que la observada.

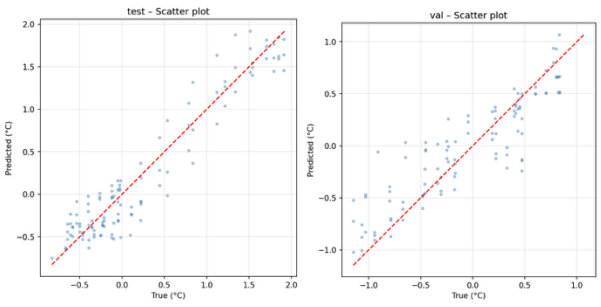


Fig. 1 Predicción frente a la observación en la prueba (izquierda) y en la validación (derecha). La línea discontinua representa el ajuste ideal; la separación aumenta en los extremos.

E. Distribución del error

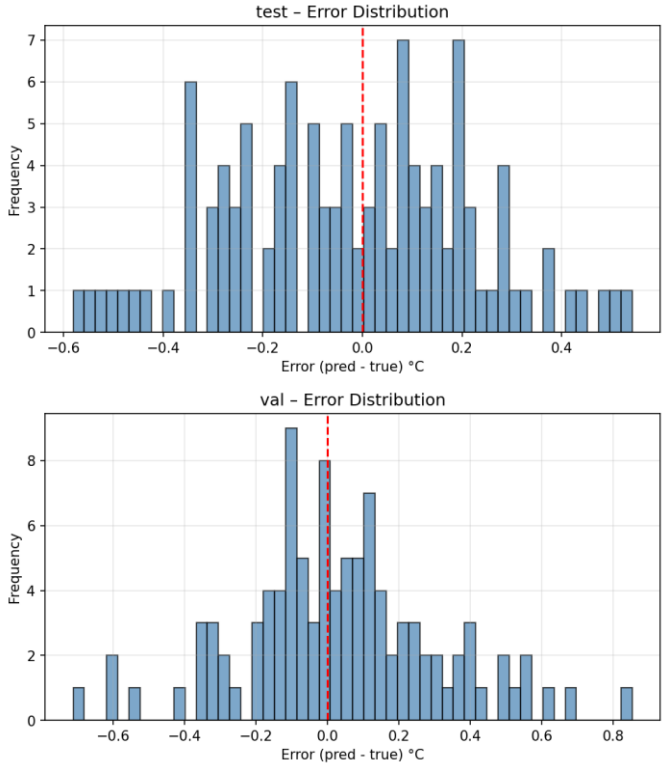
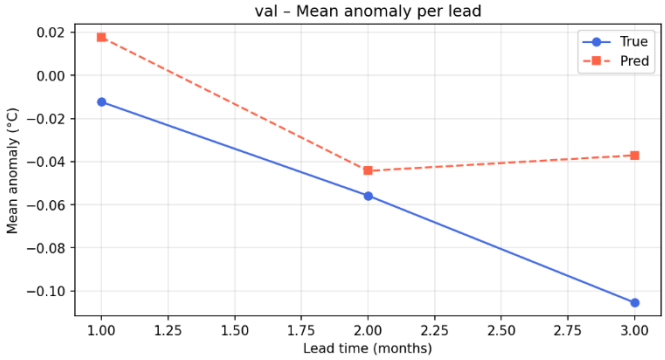


Fig. 2 Distribución del error predicho menos el observado en la prueba (arriba) y en la validación (abajo). En ambos conjuntos, la distribución está centrada cerca de cero; en la prueba, los errores se encuentran entre −0.58 y +0.54 °C.

F. Perfil medio de anomalía por lead time

La Fig. 3 resume la anomalía media por horizonte. En validación, el promedio predicho se mantiene cercano al observado, con diferencias inferiores a 0.07 °C. En prueba, la predicción media sobreestima ligeramente la anomalía en el primer mes (+0.06 °C) y la subestima en el segundo (−0.04 °C) y en el tercero (−0.10 °C). El comportamiento medio sugiere que la subestimación aumenta con el horizonte y no se debe únicamente a unos pocos casos extremos.



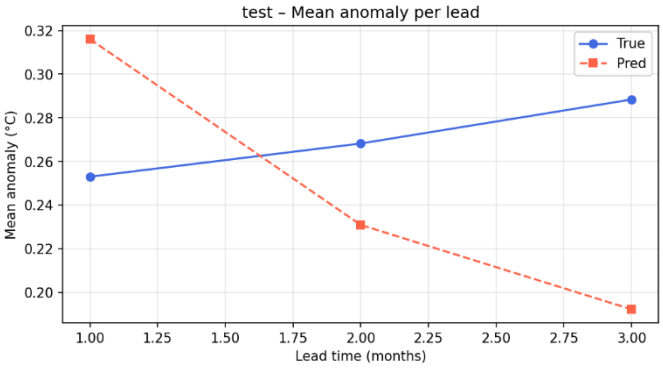


Fig. 3. Anomalía media por horizonte en validación (arriba) y en prueba (abajo). En prueba, la subestimación aumenta a medida que se alarga el horizonte.

G. Ejemplos de predicciones

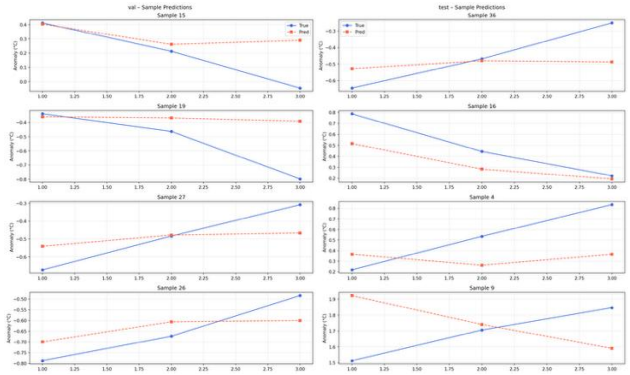


Fig. 4. Ejemplos de pronósticos individuales en la validación (izquierda) y en la prueba (derecha). El modelo sigue, en varios casos, la dirección del cambio, aunque suaviza algunos máximos y mínimos y, en otros, aplana la trayectoria de los tres meses.

H. Consistencia temporal en el lead central

La consistencia temporal se examinó en el segundo horizonte para todas las muestras. Como se observa en la Fig. 5, el modelo sigue la sucesión de calentamiento y enfriamiento, aunque atenúa los picos y presenta desfases locales. Esta lectura complementa la correlación global al mostrar periodos consecutivos en los que el error mantiene el mismo signo.



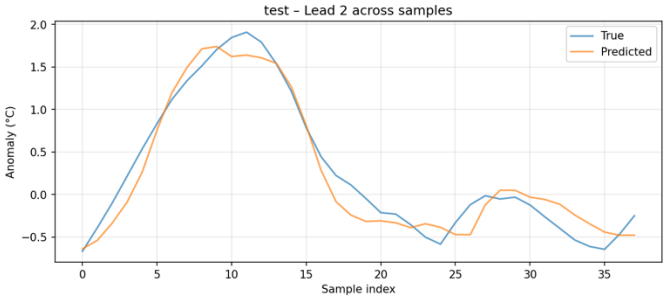


Fig. 5. Evolución del lead 2 en validación (arriba) y en prueba (abajo). La predicción sigue una estructura de baja frecuencia, con suavizado de los extremos.

I. Comparación de correlación frente a persistencia

La Fig. 6 compara la correlación por horizonte del Transformer con la persistencia. Esta última fue ligeramente superior en el primer mes (0.970 frente a 0.968). En el segundo y el tercer mes, la diferencia a favor del Transformer se hizo evidente (0.971 frente a 0.884 y 0.933 frente a 0.747), cuando mantener constante la última anomalía observada dejó de representar adecuadamente la evolución de la serie. La mejora, por tanto, se concentra en los horizontes de dos y tres meses y no es uniforme. Esta comparación describe únicamente la correlación; las diferencias de error se evalúan mediante el RMSE y mediante pruebas estadísticas en la Sección III.J.

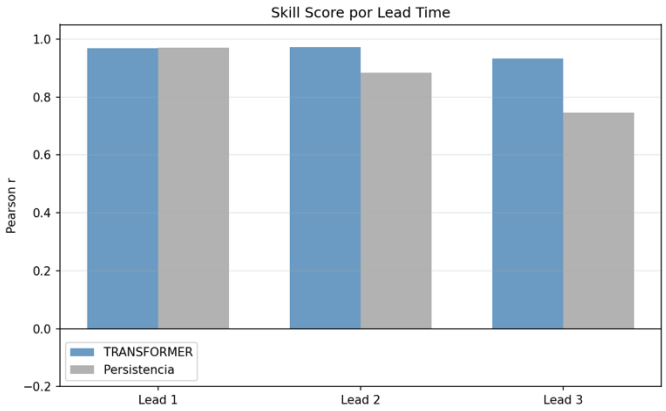


Fig. 6. Correlación de Pearson por horizonte temporal entre el Transformer y la persistencia.

J. Significancia estadística frente a persistencia

Para determinar si las diferencias de habilidad observadas en la Fig. 6 son estadísticamente significativas, se aplicaron la prueba de Diebold-Mariano [30] (H_0: igual precisión predictiva, calculada a partir del error cuadrático) y un bootstrap por bloques de la diferencia de RMSE (ΔRMSE = RMSE_Transformer − RMSE_Persistencia), ambos de forma independiente para cada lead time sobre el conjunto de prueba.

TABLA VII
SIGNIFICANCIA ESTADÍSTICA FRENTE A PERSISTENCIA (PRUEBA)

Lead	DM	p (DM)	ΔRMSE (°C)	p (bootstrap)
1	0.71	0.477	+0.025	0.486

2	−3.03	0.002	−0.201	<0.001
3	−2.80	0.005	−0.280	<0.001

Ambas pruebas coincidieron en la dirección y la significancia: las diferencias del segundo y del tercer mes fueron estadísticamente significativas ($p<0.05$), a favor del Transformer. En el lead 1, la persistencia obtuvo un RMSE ligeramente menor (+0.025 °C), pero no se pudo rechazar la hipótesis de igualdad de la habilidad predictiva.

K. Robustez frente a la semilla de inicialización

Para evaluar la estabilidad del resultado principal, el modelo se reentrenó con diez semillas distintas (Sección II.D.1), manteniendo el resto del protocolo sin cambios.

TABLA VIII
MÉTRICAS GLOBALES, MEDIA ± DESVIACIÓN ESTÁNDAR (10 SEMILLAS)

Métrica	Validación	Prueba
RMSE (°C)	0.292 ± 0.014	0.351 ± 0.050
MAE (°C)	0.234 ± 0.017	0.276 ± 0.036
R^2	0.767 ± 0.023	0.807 ± 0.054
Correlación de Pearson	0.887 ± 0.010	0.920 ± 0.024
Sesgo (°C)	−0.002 ± 0.062	−0.105 ± 0.086

Por horizonte, el RMSE de prueba promedio fue de 0.229±0.060, 0.310±0.051 y 0.465±0.076 °C para los leads 1 a 3, con correlaciones de 0.980±0.007, 0.939±0.023 y 0.851±0.052, respectivamente. La corrida individual con semilla 42, reportada en la Tabla IV, corresponde exactamente a la corrida de semilla 42 de este análisis y obtuvo un RMSE de prueba de 0.246 °C, el menor de las diez semillas y aproximadamente dos desviaciones estándar por debajo del promedio. Por ello, el promedio y la desviación estándar entre semillas constituyen la referencia representativa de la habilidad esperada, mientras que la corrida principal debe interpretarse de forma ilustrativa. Esta variabilidad, atribuible a la inicialización de pesos y al orden de los lotes (Sección II.D.1), indica que diferencias puntuales de RMSE inferiores a ~0.05 °C entre configuraciones no deben interpretarse como mejoras sustantivas sin un análisis de significancia adicional.

L. Validación con origen móvil (rolling-origin)

Para examinar si el desempeño depende específicamente del bloque de prueba 2022-presente, se evaluó el modelo con veintiún orígenes de partición sucesivos (bloques de prueba de dos años, desde 2002-01–2004-01 hasta 2022-01–2024-01), reentrenándolo en cada caso. Se excluyeron dos orígenes adicionales (2023-01–2025-01 y 2024-01–2026-01) por contener una sola muestra o ninguna, lo cual resulta insuficiente para estimar una correlación.

El RMSE agregado entre los veintiún orígenes fue de 0.458±0.106 °C y la correlación de 0.836±0.131, con una variación sustancial entre orígenes (correlación entre 0.48 y 0.95 según el bloque evaluado). Por horizonte, el RMSE promedio fue de 0.356±0.126, 0.438±0.109 y 0.553±0.100 °C para los leads 1 a 3, con correlaciones de 0.929±0.076, 0.853±0.118 y 0.730±0.197, respectivamente. Estos resultados muestran que el desempeño depende considerablemente del régimen ENSO dominante en cada bloque de prueba y que el bloque 2022-presente, evaluado en la Sección III.A, no es necesariamente representativo de otros periodos históricos.

M. Ablación del ONI

Para verificar si el ONI aporta información no redundante respecto de Niño 3.4, Niño 3 y Niño 1+2, se entrenó una variante con tres índices de entrada (sin ONI), bajo el mismo protocolo y con las mismas tres semillas (42, 123, 456). En prueba, el RMSE pasó de 0.352±0.082 °C (con ONI) a 0.368±0.033 °C (sin ONI), el R^2 de 0.800±0.087 a 0.791±0.037, y la correlación de Pearson de 0.926±0.020 a 0.937±0.011. En validación, el efecto fue igualmente pequeño (RMSE de 0.288±0.015 a 0.291±0.020 °C). La magnitud de estas diferencias, menor que la variabilidad entre semillas reportada en la Sección III.K, sugiere que la remoción del ONI no afecta sustancialmente el desempeño, lo cual es consistente con la redundancia de información entre este índice y el índice Niño 3.4, señalada en la Sección II.A.

N. Comparación con otras arquitecturas

Bajo el mismo protocolo (partición, normalización, ventana de 12 meses, tres semillas: 42, 123, 456), se entrenaron tres arquitecturas de referencia: una LSTM bidireccional (3 capas, 128 unidades, *dropout* 0.2, con atención), una CNN-LSTM (dos capas Conv1D de 128 y 256 filtros seguidas de una LSTM bidireccional de 128 unidades, *dropout* 0.2) y una TCN (cuatro bloques residuales, dilataciones 1-2-4-8, 128 filtros, kernel 3, *dropout* 0.2).

TABLA IX
COMPARACIÓN DE ARQUITECTURAS EN PRUEBA, MEDIA DE TRES SEMILLAS

Arquitectura	RMSE (°C)	MAE (°C)	R^2	Pearson r
LSTM	0.420	0.353	0.728	0.866
TCN	0.425	0.347	0.722	0.871
Persistencia	0.423	0.341	0.725	0.868
Transformer	0.352	0.272	0.800	0.926
CNN-LSTM	0.344	0.292	0.818	0.919

CNN-LSTM obtuvo el menor RMSE global (0.344 °C), ligeramente inferior al del Transformer (0.352 °C), mientras que este último alcanzó la mayor correlación de Pearson (0.926 frente a 0.919). LSTM y TCN quedaron por debajo de la persistencia en términos de RMSE y R^2. Por horizonte, CNN-LSTM fue la arquitectura más precisa en los leads 1 y 2 (RMSE=0.256 y 0.332 °C), mientras que el Transformer obtuvo la mejor correlación en los tres horizontes (0.979, 0.946 y 0.858) y el menor RMSE en el lead 3 (0.457 °C) entre los modelos de aprendizaje profundo comparados. En conjunto, el Transformer no domina de forma uniforme a las demás arquitecturas de aprendizaje profundo evaluadas bajo este protocolo, aunque en promedio obtuvo mejores métricas agregadas continuas que las de la persistencia.

O. Comparación categórica con persistencia (test de McNemar)

Se aplicaron los mismos umbrales de clasificación (±0.5 °C) a la persistencia y se compararon sus aciertos y errores de fase con los del Transformer mediante la prueba de McNemar [31] (con corrección de continuidad de Yates), por separado para cada lead time, tanto en la clasificación de fase exacta (3 clases) como en la de evento binario (neutral vs. no neutral).

En prueba, la exactitud de fase de la persistencia fue de 81.6%, 63.2% y 50.0% en los leads 1 a 3, frente a 84.2%, 81.6% y 68.4% en el Transformer. El CSI del evento binario de persistencia fue de 0.696, 0.462 y 0.321, frente a 0.714, 0.632 y 0.455 del Transformer. La diferencia resultó estadísticamente significativa en los leads 2 y 3 (χ^2=4.00, p=0.046 y χ^2=5.14, p=0.023, respectivamente), pero no en el lead 1 (χ^2=0.00, p=1.000). En la validación, no se observó ninguna diferencia significativa ($p > 0.05$ en los tres leads).

El CSI de La Niña fue consistentemente menor que el de El Niño en ambos métodos y colapsó a 0 en el tercer mes de prueba, tanto para el Transformer como para la persistencia. Este resultado coincide con el reducido número de eventos de La Niña en el bloque de prueba (desde 2022 en adelante), dominado por la fase cálida de 2023-2024, y no debe interpretarse como una limitación intrínseca del modelo frente a esa fase en particular.

IV. Discusión

El experimento muestra que cuatro índices oceánicos contienen suficiente señal para seguir la evolución de Niño 3.4 a corto plazo. Su principal interés no es competir con modelos espaciales de gran escala [7], [15], [18], sino ofrecer una representación compacta que pueda incorporarse como entrada de módulos posteriores. Esta ventaja se refiere al volumen y a la simplicidad de los datos de entrada; el estudio no midió tiempos de inferencia ni consumo de hardware.

La correlación elevada debe interpretarse junto con el RMSE, el MAE y el sesgo. El modelo reproduce la trayectoria general, pero pierde amplitud en varios extremos y la exactitud de fase desciende al 68.4% en el tercer mes. Por ello, una predicción cercana a ±0.5 °C no debería derivar automáticamente en una alerta categórica. Para ese uso, sería necesario acompañarla de intervalos o de probabilidades de excedencia, y de información oceánico-atmosférica adicional.

La tendencia a un sesgo negativo, observada en nueve de las diez semillas y en el tercer mes de la corrida principal, puede deberse a varios factores: un cambio de distribución respecto del entrenamiento, la composición particular de los eventos recientes o el suavizado habitual de las salidas deterministas. La validación con origen móvil mostró que el desempeño varía sustancialmente entre los bloques de prueba. Este resultado indica sensibilidad al régimen ENSO evaluado, aunque los experimentos realizados no permiten separar de forma concluyente el efecto del cambio de distribución de un posible sesgo asociado a la arquitectura.

Las limitaciones más importantes se derivan de la representación de entrada. Los índices escalares no describen explícitamente el viento, la profundidad de la termoclina, el contenido de calor, la salinidad ni las teleconexiones espaciales. La ablación del ONI produjo cambios pequeños: el RMSE de prueba aumentó en menos de 0.02 °C al excluir este predictor, una diferencia inferior a la variabilidad observada entre semillas. Este resultado sugiere que su contribución incremental es limitada. El bloque de prueba es reciente y relativamente corto; la comparación bajo el mismo protocolo con LSTM, CNN-LSTM y TCN mostró que el Transformer no domina de forma uniforme frente a otras arquitecturas de aprendizaje profundo — CNN-LSTM obtuvo un RMSE global ligeramente menor —, aunque sí frente a la persistencia en el desempeño agregado. No se compararon variantes modernas de Transformer, como TFT, Informer, FEDformer o PatchTST, bajo el mismo protocolo.

Como continuación de este estudio, se plantea evaluar ventanas históricas de 6, 12, 18 y 24 meses, medir con mayor granularidad el aporte individual de cada predictor y generar salidas probabilísticas mediante cuantiles o ensambles. La incorporación de contenido de calor, viento y salinidad permitiría estudiar si una entrada físicamente más rica reduce la barrera de primavera y el sesgo observado [18], [19].

La clasificación utilizada en este trabajo distingue entre meses cálidos, fríos o neutrales en función de la anomalía mensual de Niño 3.4. Este criterio es útil para evaluar la discriminación inmediata del modelo, pero no equivale a la definición operacional de un episodio de ENSO, que considera promedios móviles y la persistencia a lo largo de varias temporadas superpuestas. En consecuencia, el CSI y las métricas asociadas describen la identificación de meses no neutrales en este conjunto, no la validación de un sistema oficial de alerta. Cabe destacar que el CSI de La Niña fue sistemáticamente menor que el de El Niño y colapsó en el tercer mes de prueba, un patrón más asociado al reducido número de eventos fríos en el bloque de prueba que a una limitación estructural del modelo en esa fase.

La separación cronológica, el ajuste de los escaladores únicamente con datos de entrenamiento y el aislamiento del conjunto de prueba reducen el riesgo de fuga de información. Aun así, la arquitectura y sus hiperparámetros se seleccionaron con un único periodo de validación, 2018-2021, y podrían haberse adaptado a ese régimen. Las ventanas mensuales también se superponen; esta dependencia es propia de un pronóstico actualizado cada mes, pero reduce el tamaño efectivo de la muestra y puede hacer que las diferencias pequeñas parezcan más estables.

La generalización de los resultados está condicionada por la cobertura temporal, el producto de SST utilizado y el uso exclusivo de los índices del Pacífico oriental y central. El desempeño puede cambiar con otra climatología base, revisiones del producto o eventos costeros que no se reflejen de la misma manera en Niño 3.4. Los tests de Diebold-Mariano y McNemar confirmaron que la ventaja del Transformer frente a

la persistencia es estadísticamente significativa en el segundo y el tercer mes, tanto en la métrica continua como en la clasificación categórica; no así en el primer mes, donde ninguna diferencia fue significativa. Por ello, la correlación elevada observada en el conjunto agregado respalda la existencia de capacidad predictiva en el periodo analizado, pero la ventaja estadísticamente significativa frente a la persistencia se concentra en los horizontes de dos y tres meses.

V. Conclusiones

En conclusión, se entrenó un Transformer Encoder con 12 meses de datos de cuatro índices oceánicos para pronosticar el índice Niño 3.4 con 1, 2 y 3 meses de anticipación. Con una separación estrictamente cronológica, la corrida principal determinista alcanzó en prueba un RMSE de 0.2456 °C, un R^2 de 0.9072 y una correlación de 0.9539. Esta corrida fue la de menor error entre las 10 semillas evaluadas, cuyo RMSE promedio de 0.351 ± 0.050 °C representa el desempeño esperado del modelo. En el tercer mes, la correlación fue de 0.933, y el CSI en los meses no neutrales alcanzó 0.597, con 0.892 para El Niño y 0.160 para La Niña.

Una prueba de Diebold-Mariano y un bootstrap por bloques mostraron que la ventaja del Transformer frente a la persistencia fue estadísticamente significativa en el segundo y el tercer mes, pero no en el primero; una prueba de McNemar sobre la clasificación categórica mostró el mismo patrón. La comparación bajo el mismo protocolo con LSTM, CNN-LSTM y TCN mostró que el Transformer no domina de forma uniforme frente a otras arquitecturas de aprendizaje profundo —CNN-LSTM obtuvo un RMSE global ligeramente menor (0.344 vs. 0.352 °C)—, aunque sí lo hizo en términos de desempeño agregado. La validación con origen móvil confirmó que el desempeño depende considerablemente del régimen ENSO evaluado (RMSE de 0.458±0.106 °C y correlación de 0.836±0.131 entre veintiún orígenes), y la ablación del ONI indicó que este predictor es en gran medida redundante con Niño 3.4.

En conjunto, los resultados indican que una arquitectura compacta basada en índices conserva la capacidad predictiva a corto plazo. Frente a la persistencia, la mejora fue estadísticamente significativa en el segundo y el tercer mes, tanto en el error continuo como en la clasificación categórica. La tendencia a subestimar las anomalías entre semillas, la pérdida de exactitud categórica en el tercer mes, la menor capacidad de detección de La Niña frente a El Niño y la ausencia de una medida de incertidumbre son las principales restricciones. La siguiente versión experimental deberá incorporar predictores físicos adicionales, realizar comparaciones con variantes modernas de Transformer (TFT, Informer, PatchTST) y generar salidas probabilísticas antes de evaluar su integración en un sistema de apoyo a la gestión del riesgo.

Agradecimiento

Se agradece a la Universidad Católica de Santa María y al Vicerrectorado de Investigación por su apoyo al desarrollo del proyecto de investigación con Resolución 31951-R-2025 “Sistema modular de alerta temprana para deslizamientos inducidos por El Niño en el norte del Perú: Predicción escalonada del ENSO, precipitaciones extremas y susceptibilidad mediante aprendizaje automático”, ahora P PI-2025-1-016.

Referencias

[1] K. E. Trenberth, “The Definition of El Niño,” *Bulletin of the American Meteorological Society*, vol. 78, no. 12, 1997, doi: https://doi.org/10.1175/1520-0477(1997)078<2771:TDOENO>2.0.CO;2.

[2] M. J. McPhaden, S. E. Zebiak, and M. H. Glantz, “ENSO as an Integrating Concept in Earth Science,” *Science*, vol. 314, no. 5806, pp. 1740–1745, Dec. 2006, doi: 10.1126/science.1132588.

[3] D. Chen, M. A. Cane, A. Kaplan, S. E. Zebiak, and D. Huang, “Predictability of El Niño over the past 148 years,” *Nature*, vol. 428, no. 6984, pp. 733–736, Apr. 2004, doi: 10.1038/nature02439.

[4] A. G. Barnston, M. K. Tippett, M. L. L’Heureux, S. Li, and D. G. DeWitt, “Skill of Real-Time Seasonal ENSO Model Predictions during 2002–11: Is Our Capability Increasing?,” *Bulletin of the American Meteorological Society*, vol. 93, no. 5, 2012, doi: 10.1175/BAMS-D-11-00111.1.

[5] R. Emberson, D. Kirschbaum, and T. Stanley, “Global connections between El Nino and landslide impacts,” *Nat Commun*, vol. 12, no. 1, p. 2262, Apr. 2021, doi: 10.1038/s41467-021-22398-4.

[6] M. Yglesias-González *et al.*, “Reflections on the impact and response to the Peruvian 2017 Coastal El Niño event: Looking to the past to prepare for the future,” *PLOS ONE*, vol. 18, no. 9, p. e0290767, Sep. 2023, doi: 10.1371/journal.pone.0290767.

[7] Y.-G. Ham, J.-H. Kim, and J.-J. Luo, “Deep learning for multi-year ENSO forecasts,” *Nature*, vol. 573, no. 7775, pp. 568–572, Sep. 2019, doi: 10.1038/s41586-019-1559-7.

[8] Y. Wang, H. A. Karimi, and X. Jia, “Deep Learning Model for ENSO Forecasting Using Multiple-Scale Spatiotemporal Information,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–10, 2025, doi: 10.1109/TGRS.2025.3529322.

[9] J. Yan, L. Mu, L. Wang, R. Ranjan, and A. Y. Zomaya, “Temporal Convolutional Networks for the Advance Prediction of ENSO,” *Sci Rep*, vol. 10, no. 1, p. 8055, May 2020, doi: 10.1038/s41598-020-65070-5.

[10] P. J. Petersik and H. A. Dijkstra, “Probabilistic Forecasting of El Niño Using Neural Network Models,” *Geophysical Research Letters*, vol. 47, no. 6, p. e2019GL086423, 2020, doi: 10.1029/2019GL086423.

[11] Y.-G. Ham, J.-H. Kim, E.-S. Kim, and K.-W. On, “Unified deep learning model for El Niño/Southern Oscillation forecasts by incorporating seasonality in climate data,” *Science Bulletin*, vol. 66, no. 13, pp. 1358–1366, Jul. 2021, doi: 10.1016/j.scib.2021.03.009.

[12] J. Hu, B. Weng, T. Huang, J. Gao, F. Ye, and L. You, “Deep Residual Convolutional Neural Network Combining Dropout and Transfer Learning for ENSO Forecasting,” *Geophysical Research Letters*, vol. 48, no. 24, p. e2021GL093531, 2021, doi: 10.1029/2021GL093531.

[13] C. Song, X. Zhang, X. Chen, H. Jiang, L. Zhang, and Y. Huang, “A deep residual intelligent model for ENSO prediction by incorporating coupled model forecast data,” *Acta Oceanol. Sin.*, vol. 44, no. 8, pp. 133–142, Aug. 2025, doi: 10.1007/s13131-025-2510-4.

[14] B. Mu, B. Qin, and S. Yuan, “ENSO-GTC: ENSO Deep Learning Forecast Model With a Global Spatial-Temporal Teleconnection Coupler,” *Journal of Advances in Modeling Earth Systems*, vol. 14, no. 12, p. e2022MS003132, 2022, doi: 10.1029/2022MS003132.

[15] L. Zhou and R.-H. Zhang, “A self-attention–based neural network for three-dimensional multivariate modeling and its skillful ENSO predictions,” *Sci. Adv.*, vol. 9, no. 10, p. eadf2827, Mar. 2023, doi: 10.1126/sciadv.adf2827.

[16] C. Gao, L. Zhou, and R.-H. Zhang, “A Transformer-Based Deep Learning Model for Successful Predictions of the 2021 Second-Year La Niña

Condition," *Geophysical Research Letters*, vol. 50, no. 12, p. e2023GL104034, 2023, doi: 10.1029/2023GL104034.
[17] H. Wang, S. Hu, and X. Li, "An Interpretable Deep Learning ENSO Forecasting Model," *Ocean-Land-Atmosphere Research*, vol. 2, p. 0012, Jun. 2023, doi: 10.34133/olar.0012.
[18] B. Mu, Y. Cui, S. Yuan, and B. Qin, "Incorporating heat budget dynamics in a Transformer-based deep learning model for skillful ENSO prediction," *npj Clim Atmos Sci*, vol. 7, no. 1, p. 208, Sep. 2024, doi: 10.1038/s41612-024-00741-y.
[19] H. Wang, S. Hu, C. Guan, and X. Li, "The role of sea surface salinity in ENSO forecasting in the 21st century," *npj Clim Atmos Sci*, vol. 7, no. 1, p. 206, Sep. 2024, doi: 10.1038/s41612-024-00763-6.
[20] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Aug. 02, 2026. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
[21] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal Fusion Transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, Oct. 2021, doi: 10.1016/j.ijforecast.2021.03.012.
[22] H. Zhou *et al.*, "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, pp. 11106–11115, May 2021, doi: 10.1609/aaai.v35i12.17325.
[23] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2021, pp. 22419–22430. Accessed: Aug. 02, 2026. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/hash/bcc0d400288793e8bdcd7c19a8ac0c2b-Abstract.html
[24] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, "FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting," in *Proceedings of the 39th International Conference on Machine Learning*, PMLR, Jun. 2022, pp. 27268–27286. Accessed: Aug. 02, 2026. [Online]. Available: https://proceedings.mlr.press/v162/zhou22g.html
[25] R. Magallanes-Quintanar, C. E. Galván-Tejada, J. I. Galván-Tejada, S. de J. Méndez-Gallegos, and A. García-Domínguez, "Standardized Precipitation Index Forecasting Comparison Using Transformer Models," *Forecasting*, vol. 8, no. 3, p. 44, Jun. 2026, doi: 10.3390/forecast8030044.
[26] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A Time Series is Worth 64 Words: Long-term Forecasting with Transformers," in *International Conference on Learning Representations (ICLR 2023)*, 2023. Accessed: Sep. 27, 2026. [Online]. Available: https://openreview.net/forum?id=Jbdc0vTOcol
[27] B. Huang *et al.*, "Improvements of the Daily Optimum Interpolation Sea Surface Temperature (DOISST) Version 2.1," *Journal of Climate*, vol. 34, no. 8, pp. 2923–2939, Mar. 2021, doi: 10.1175/JCLI-D-20-0166.1.
[28] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer Normalization," arXiv.org. [Online]. Available: https://arxiv.org/abs/1607.06450v1
[29] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations (ICLR 2019)*, Sep. 2018. doi: 10.48550/arXiv.1711.05101.
[30] F. X. Diebold and R. S. Mariano, "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, vol. 13, no. 3, pp. 253–263, 1995, doi: 10.2307/1392185.
[31] Q. McNemar, "Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, Jun. 1947, doi: 10.1007/BF02295996.