Article

Combined dynamical-deep learning ENSO forecasts

Yipeng Chen [1,9], Yishuai Jin [1,2,9] ✉, Zhengyu Liu [3] ✉, Xingchen Shen [4], Xianyao Chen [1,5], Xiaopei Lin [1,5], Rong-Hua Zhang [6], Jing-Jia Luo [6], Wenjun Zhang [6], Wansuo Duan [7], Fei Zheng [7], Michael J. McPhaden [8] & Lu Zhou [6]

Improving the prediction skill of El Niño-Southern Oscillation (ENSO) is of critical importance for society. Over the past half-century, significant improvements have been made in ENSO prediction. Recent studies have shown that deep learning (DL) models can substantially improve the prediction skill of ENSO compared to individual dynamical models. However, effectively integrating the strengths of both DL and dynamical models to further improve ENSO prediction skill remains a critical topic for in-depth investigations. Here, we show that these DL forecasts, including those using the Convolutional Neural Networks and 3D-Geoformer, offer comparable ENSO forecast skill to dynamical forecasts that are based on the dynamic-model mean. More importantly, we introduce a combined dynamical-DL forecast, an approach that integrates DL forecasts with dynamical model forecasts. Two distinct combined dynamical-DL strategies are proposed, both of which significantly outperform individual DL or dynamical forecasts. Our findings suggest the skill of ENSO prediction can be further improved for a range of lead times, with potentially far-reaching implications for climate forecasting.

El Niño-Southern Oscillation (ENSO) is the strongest interannual climate variability on our planet[1,2]. It exerts strong impacts on regional climate and society worldwide[3–5] through atmospheric teleconnection[6]. The prediction of ENSO has challenged the climate community for the last half century. One major difficulty is the ENSO spring predictability barrier (SPB), referring to a rapid decrease in prediction skill during the boreal spring regardless of the different initial months[7–11]. Intensive efforts have been made in improving ENSO prediction skill in two approaches: dynamical modeling and statistical modeling[12–17]. The dynamical models, based primarily on the physical equations of the ocean–atmosphere system, range from simplified physics[18,19] to state-of-the-art comprehensive fully

coupled general circulation models[20]. The statistical models, which have been developed primarily using historical observational datasets, employ statistical models from various forms of regression[21–24] to nonlinear machine learning strategies[13,25]. However, both approaches have deficiencies. Dynamical prediction still suffers from problems in model systematic error and initialization, while statistical methods are limited by the length of observational data and the nature of the statistical model. As such, for a long period before, dynamical forecasts and statistical forecasts tend to have comparable skills[13,24,26].

Recently, more advanced deep learning (DL) models have been developed for ENSO prediction[27]. In contrast to traditional statistical

[1]Frontier Science Center for Deep Ocean Multispheres and Earth System (FDOMES) and Physical Oceanography Laboratory, Ocean University of China, Qingdao, China. [2]SANYA Oceanographic Laboratory, Sanya, China. [3]Atmospheric Science Program, Department of Geography, The Ohio State University, Columbus, OH, USA. [4]National Key Laboratory of Space Integrated Information System, Institute of Software, Chinese Academy of Sciences, Beijing, China. [5]Laoshan Laboratory, Qingdao, China. [6]State Key Laboratory of Climate System Prediction and Risk Management, Nanjing University of Information Science and Technology, Nanjing, China. [7]State Key Laboratory of Earth System Numerical Modeling and Application, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China. [8]National Oceanic and Atmospheric Administration/Pacific Marine Environmental Laboratory, Seattle, WA, USA. [9]These authors contributed equally: Yipeng Chen, Yishuai Jin. ✉e-mail: jinyishuai@126.com; liu.7022@osu.edu

models that are mainly trained on the limited historical observation only[25,28], most DL models are trained on a dramatically larger data set from the simulations of dozens of state-of-the-art dynamical climate models and have demonstrated significantly enhanced prediction skill over most individual dynamical models[29–32]. Meanwhile, a recent study suggests that an extended nonlinear recharge oscillator model, a traditional statistical model interacting with climate dynamics, performs better than several dynamical models and is comparable to most DL models[33]. This implies that the integration of dynamical knowledge and statistical algorithms plays a unique role in ENSO forecasting. However, how to effectively combine the two to further enhance ENSO prediction remains an important topic for future research.

Here, we demonstrate that ENSO prediction, including skill improvement through the SPB, can be significantly enhanced across various lead times by integrating DL forecasts with dynamical model forecasts. This approach, referred to as the combined dynamical-DL forecast—or simply, the Dynamical-DL Forecast—offers substantial predictive benefits.

Results

DL forecasts vs dynamical forecasts

Previous studies showed that ENSO prediction skill is lower in most individual dynamical models of the North American Multi-Model Ensemble (NMME) than in recent DL models[29–31]. However, here, we show that the dynamic-model mean of the forecast skill of NMME models is comparable to that of DL forecasts. We reproduce the ENSO forecasts in two DL models (Convolutional Neural Networks (CNN) and 3D-Geoformer, see "Methods") that are trained on Coupled Model Intercomparison Project phase 6 (CMIP6) model simulations (see "Methods"), as in previous studies[29,31]. Consistent with these studies,

the useful forecast skill threshold of 0.5, as represented by the Anomaly Correlation Coefficients (ACC, see "Methods") of the year-round forecast Niño3.4 index of the two DL models, is extended to about 16−18 lead months (Fig. 1a, thick red and pink lines). These DL-derived year-round ACCs (Fig. 1a) and spring ACCs (Fig. 1b), when the prediction is made during March, are significantly higher than 75% and 62.5% of individual NMME models, respectively. However, the forecast skill of the dynamic-model mean (black line in Fig. 1a, b) is comparable with that of the two DL forecasts for both the year-round forecast (red and pink lines in Fig. 1a) and for crossing the SPB (Fig. 1b−e). Similar features can also be identified when we use root mean square error (RMSE; see "Methods"; the DL-derived spring RMSEs are significantly lower than 62.5% of the NMME models) to assess ENSO predictability (Supplementary Fig. 1). Moreover, an examination of the spatial distribution of springtime ACC reveals that the dynamic-model mean ACC is higher than that in the DL model of 3D-Geoformer (Supplementary Fig. 2, note also that CNN does not predict spatial distribution), especially for the central-eastern Pacific, although the DL ACCs become somewhat higher than that of the dynamic-model mean at longer leads (Supplementary Fig. 3).

The enhanced prediction skill of the dynamical forecast of the dynamic model mean over more than 75% of individual models is well known for weather and climate forecasts, as in the case of ENSO forecasts in the NMME models[34]. The enhanced prediction skill is contributed to, partly, by the suppression of model biases in the dynamic-model mean[35]. Indeed, our further analysis of the forecast errors across the NMME models shows that a larger forecast error and lower forecast skill tend to correlate with a greater tropical bias of the model climatology. The dynamic-model mean shows the smallest tropical bias and, in turn, the highest forecast skill, especially for
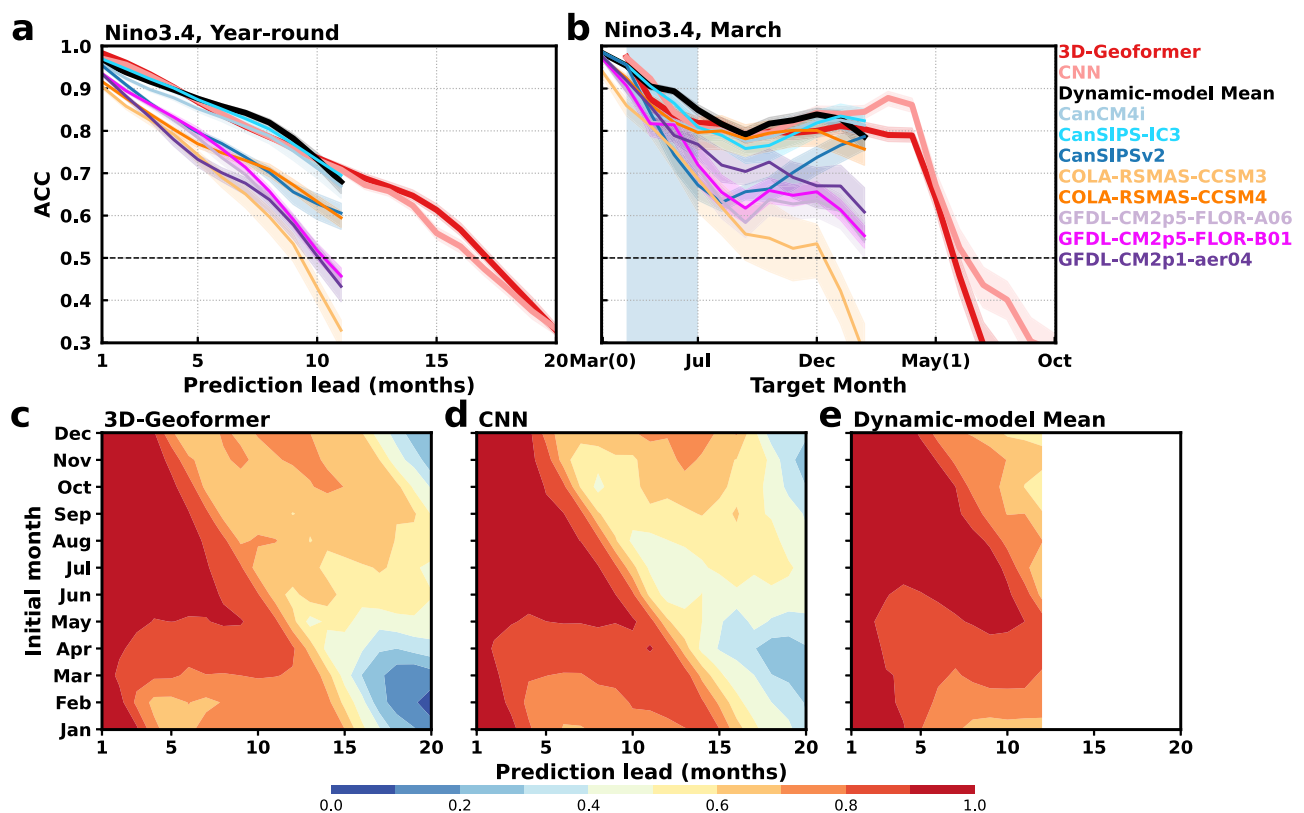
Fig. 1 | State of ENSO prediction skill for DL models and dynamical models.
a, b The ACC of DL models (thick red and pink lines) and dynamical models (black and other colors lines) from 1982 to 2018 for year-round (**a**) and for the initial month in March (**b**). **c**−**e** The seasonality prediction skill with shading of ACC during 1982-2018 for 3D-Geoformer (**c**), CNN (**d**), and dynamic-model mean (**e**) forecasts, respectively. The shading around the lines denotes the 95% confidence interval, based on the bootstrap method (see "Methods"). Note that the intervals of DL models are relatively small, making them less visually apparent in the figures.

Fig. 2 | Strategy 1, dynamical-DL based on ensemble mean forecast. a The ACC of 3D-Geoformer (red), CNN (pink), dynamical-model mean (black), and strategy 1 (blue) for year-round. **b** Same as (**a**), except for the initial month in March. **c** The seasonality of the prediction skill with shading of ACC of Strategy 1 during 1982–2018. The shading around the lines denotes the 95% confidence interval, based on the bootstrap method. Note that the intervals of DL models are relatively small, making them less visually apparent in the figures.
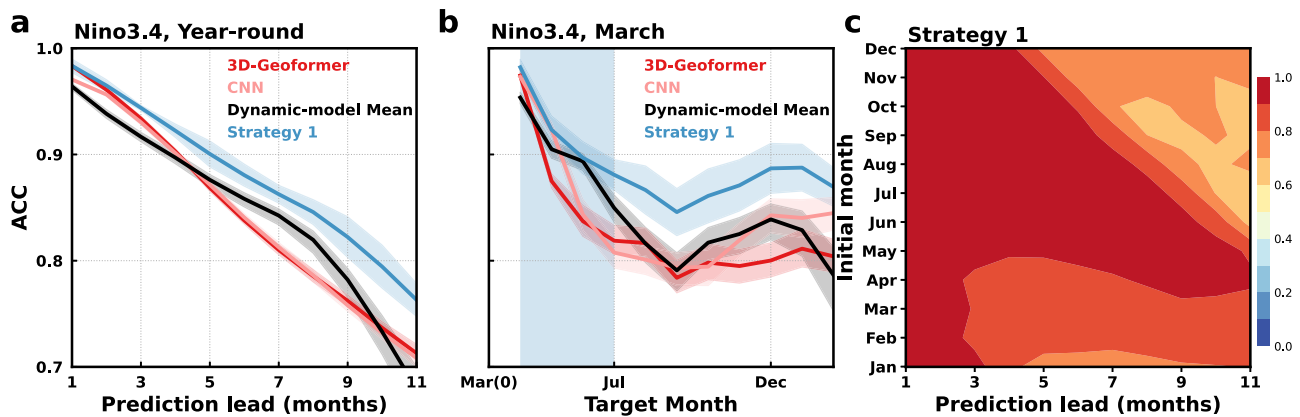
crossing the SPB (Supplementary Discussion 1 and Supplementary Figs. 4–9). Note that we define the model tropical bias as the mean of the absolute climatology bias of the region 150° E–80° W, 5° S–5° N, (black box in Supplementary Fig. 5) between observation and model for the corresponding targeted month during 1982–2010.

It is interesting to see here that 6 out of 8 models are inferior to an advanced DL model forecast, but the dynamic-model mean forecast can achieve forecast skill comparable to the advanced DL models. A single dynamic model forecast does have the advantages of advanced dynamics and a sophisticated initialization strategy. However, it also has disadvantages in comparison with a single DL model, as well as the ensemble mean dynamical forecast. Aside from a larger model tropical bias than the dynamic-model mean as discussed above, a single model is limited in predictability information only to itself, while a DL model, or a multi-dynamical model ensemble mean forecast, derives predictability information from multiple dynamic models. Moreover, these sensitivity experiments (i.e., transfer learning) show that the success of the two DL forecast models in this study is due mainly to the training from the information in the large amount of CMIP6 simulations (historical simulations of 31 models, Supplementary Table 1), with little contribution from observations (Supplementary Fig. 10). Note here we do not perform transfer learning for the 3D-Geoformer, because it leads to degradation of ENSO prediction[31].

Further improving the ENSO forecast with combined dynamical-DL forecasts

The comparable skills of the dynamic-model mean forecast with the two DL forecasts suggest a distinct value of the dynamic-model mean forecast independent of the DL forecasts. This leads us to hypothesize that ENSO prediction skill can be further enhanced beyond a single DL forecast or the dynamical multi-model mean forecast if the two types of forecasts are combined into a combined dynamical-DL forecast. The first dynamical-DL strategy, referred to as Strategy 1 here (blue lines in Fig. 2, see "Methods"), is the simple average of the forecasts of the NMME dynamic-model ensemble mean and the two DL forecasts from CNN and 3D-Geoformer. This simplest of weighting strategies illustrates our general idea that climate forecasts can be further improved by combining both dynamical and DL forecasts. The optimal weighting strategy is an interesting question for future study. The prediction skill (ACC) of Strategy 1 is indeed increased significantly beyond the skills of either the DL models or the dynamical multi-model mean, for year-round forecasts except for 7–8 months lead months and for forecasts across the SPB beyond 4 lead months (Fig. 2). For the different initial months predictions, out of 132 total targets (11 lead months × 12 initial

months), "Strategy 1" achieves statistically significant ACC improvements in 47 targets (35.6%) over the 3D-Geoformer, 68 targets (51.5%) over the CNN, and 67 targets (50.8%) over the dynamical-model mean (Supplementary Fig. 11). Moreover, for RMSE, Strategy 1 shows relatively modest improvement (Supplementary Fig. 12). Specifically, Strategy 1 decreases significantly over the CNN or 3D-Geoformer beyond 3 lead months and over the dynamical-model mead except for 7–9 lead months for year-round forecasts (Supplementary Fig. 12a). For forecasts across the SPB, Strategy 1 decreases significantly over the CNN or 3D-Geoformer at 2–7 lead months and over the dynamical-model mean at 4–6 months (Supplementary Fig. 12b). Although the lead time of the available dynamical forecast and, in turn, of Strategy 1, is limited to one year, it is conceivable that the improved forecast skill should be extended to well beyond one year.

The second dynamical-DL forecast strategy, referred to as Strategy 2, is designed to improve the dynamical forecast of both single and multiple models, in contrast to Strategy 1, which only applies when there are forecasts of multiple models. In Strategy 2, a DL forecast is used as the "First-Guess" that is used to select a subset of initial conditions for a single dynamical model to perform ensemble forecast[36,37] (see "Methods"). This approach not only improves the initial condition of the single dynamical model from the original set of a large number of initial conditions but also enhances prediction skill across a range of lead times. The prediction skill of the DL models is higher than that of 6 out of 8 dynamical models, and maybe the DL models could help pick more accurate initial conditions for the single dynamical models. The subset thus selected represents a set of more optimal initial conditions with which the dynamical-DL forecast should be improved. In addition, the enhanced prediction skill is achieved by selecting ensemble members that closely align with the superior performance of DL models. This is indeed the case, as seen in the year-round forecast ACC of Strategy 2. For 5 out of 8 individual models in NMME, Strategy 2 leads to a higher ACC forecast skill than the ensemble mean dynamical forecast skill (significantly higher beyond the eight lead months) of this model alone using the original large set of initial conditions (Fig. 3a–h, thick dot lines for 3D-Geoformer and thick dashed lines for CNN against solid lines for all initial condition ensemble mean). The forecast ACC of Strategy 2 is, naturally, still below the upper limit that is produced by the hindcast that uses the "Truth" (i.e., observation; see "Methods") to select the subset of the initial condition for the individual models (thin dash-dot lines in Fig. 3). For 5 out of 8 individual NMME models at the longer lead times, similar ACC of Strategy 2 can be seen in crossing the SPB as shown in the forecast initialized from March (Supplementary Fig. 13). It is worth noting that similar
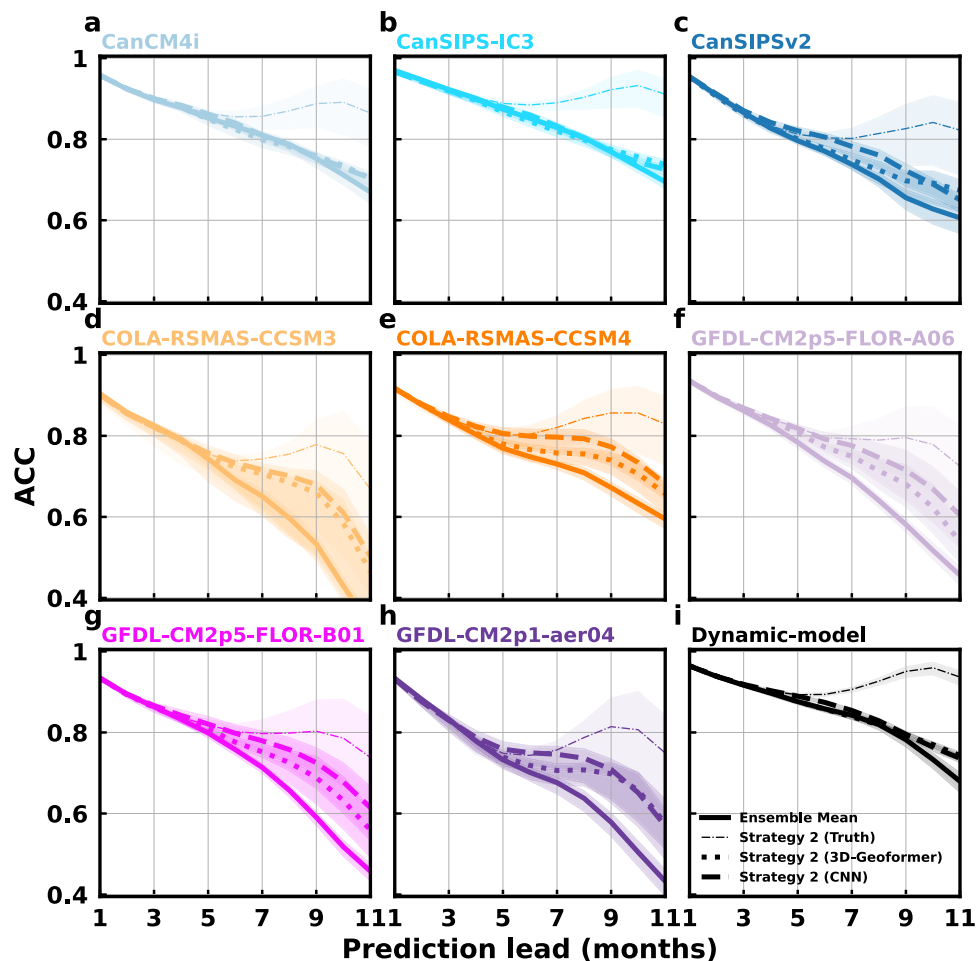
Fig. 3 | Strategy 2, dynamical-DL based on "First-Guess". The colorful and black lines indicate the prediction skill of each dynamical model (**a**–**h**) and the dynamic-model mean (**i**) for the year-round ACC during 1982–2018. The solid lines indicate the prediction skill of the all-ensemble mean. The thin dashed-dot lines indicate the prediction skill based on Strategy 2, where the "Truth" value (i.e., the observation) is an indicator (sub-ensemble members are 50% of all-ensemble members). And the dashed (CNN) and dot (3D-Geoformer) lines indicate the prediction skill based on Strategy 2, where the results of DL models are used as indicators (sub-ensemble members are 50% of all-ensemble members). The shading around the lines denotes the 95% confidence interval, based on the bootstrap method.

prediction improvement in year-round (Supplementary Fig. 14) and initialized in March (Supplementary Fig. 15) of Strategy 2 can be seen in RMSE.

Strategy 2 can also be applied to the dynamic-model mean and improves the forecast over the pure dynamical multi-model mean. Throughout the year, Strategy 2 significantly improves forecasts at the 10–11 lead months for both ACC (Fig. 3i) and RMSE (Supplementary Fig. 14i). In March-initiated forecasts, Strategy 2 (CNN) outperforms Strategy 2 (3D-Geoformer), significantly improving forecast skill between 7-11 lead months (Supplementary Fig. 15i). But for the forecast skill of CanCM4i, CanSIPS-IC3 and dynamic-model mean which using 3D-Geoformer as an indicator (Fig. 3a, b, c, and i), the improvement is not as significant (smaller than 0.05) as for other individual model because the similar forecast skill between them and DL models (Fig. 1a). Notably, by using the "First-Guess" method, the final prediction skills of ENSO in the dynamical models (5 out of 8) can be further improved by Strategy 2, particularly beyond 8 months. Specifically, when we use CNN to select a subset of the ensemble forecast of a single dynamical model (e.g., COLA-RSMAS-CCSM4 in Fig. 3e), the prediction skill can be increased from 0.67 to 0.77 at about 9 lead months, suggesting that the ability of the dynamical model in predicting ENSO is underestimated. Note that, for short lead months, the results of Strategy 2 are not sensitive to the threshold for the selection of the subset (Supplementary Fig. 16a, b) at shorter lead times (<6 months),

but it exhibits significant sensitivity at longer lead times. To avoid selecting too many or too few samples, we use the 50% threshold.

Overall, both combined dynamical-DL strategies demonstrate improved forecast skill compared to either the dynamic-model mean or individual DL models. This improvement is observed across various lead times, both for year-round ACC (e.g., Supplementary Fig. 17) and for ACC when the initial month is in spring (e.g., Supplementary Fig. 18). However, for year-round forecasts, the improvement of Strategy 2 over the dynamic-model mean is not significant at lead times of 1–4 months and 8–9 months.

Discussion

In spite of the higher ENSO prediction skill of DL models than 75% of single dynamical models, the skill of DL models is comparable to the dynamic-model mean in NMME for both the year-round forecast and the forecast through the SPB. This finding suggests that both dynamical forecasts and DL forecasts are invaluable for further improvement of the forecast skill, and an optimal forecast should utilize both dynamical and DL methods. Here we proposed a simple, yet effective strategy, the combined dynamical-DL forecast, and show it improves the ENSO forecast. In particular, Strategy 1 improves ACC over either the DL models or the dynamic-model mean for year-round forecasts, except for 7–9 months lead months and across the SPB beyond 4 lead months. Strategy 2 improves ACC over 5 out of 8 dynamical models for

year-round and across the SPB beyond 8 lead months. Furthermore, with improved climate models and more independent forecast strategies for both dynamical and DL models, our strategy opens the door for further improvement of the ENSO forecast skill in the future.

It is important to note that while the two proposed strategies enhance ENSO forecasting, they have certain limitations in real-time forecasting. As shown in Supplementary Fig. 19, Strategy 2 improves the prediction of ENSO events in 1997/98, 1998/99, 2007/08, 2008/09, 2009/10, and 2015/16 by correcting forecast results for 1, 6, 1, 5, 2, and 5 NMME models, respectively (out of 8 models in total). However, DL models systematically underestimate peak intensities of ENSO events, a well-documented issue in previous studies[29,30]. Although ref. 38. Improved ENSO peak intensity prediction by modifying seasonal-independent parameters and loss functions in CNN, the underestimation issue persists. Fundamentally, this stems from two key factors: (1) the network architecture may overly rely on a normal distribution assumption, leading to excessively smooth outputs, and (2) the limited number of extreme ENSO events in the training dataset reduces the DL model's ability to capture peak intensities, potentially constraining its operational forecasting capability.

Additionally, both proposed strategies rely on forecasts from dynamical models and DL models, posing a common challenge in real-time forecasting[39,40]. The reliance of DL models on gridded reanalysis datasets limits their real-time forecasting capability, which is limited by the update latency of reanalysis datasets; for instance, ORAS5 reanalysis datasets typically lag by half a month, preventing immediate forecasting. In a real-world scenario, on March 15, only February's reanalysis datasets would be available, allowing for forecasts initialized in February. Compared to forecasts initialized in March (Supplementary Fig. 19), Strategy 2 initialized in February improves predictions for 1, 5, 4, 5, 3, and 5 NMME models (out of 8, Supplementary Fig. 20). Future research should address this by incorporating such as ref. 38, which CNN modifications while integrating physical information into DL models to mitigate the sample size limitation. Additionally, developing DL-based climate forecasting models that utilize scattered observational data instead of reanalysis grids could reduce dependence on the update latency, further optimizing the proposed strategies and other DL models, improving their ability to simulate extreme ENSO events and real-time forecasting.

Methods
Reanalysis and model outputs
Monthly sea surface temperature (SST) and sea surface height (SSH) fields from 1982 to 2018 are obtained from the European Centre for Medium-Range Forecasts Ocean Reanalysis System 5 (ORAS5), which is the validation of ENSO prediction by DL and dynamical models. We use the monthly SST and SSH fields of 31 CMIP6 (Supplementary Table 1) in historical simulations from 1900 to 2014 as the training data of DL models. For the CNN model, we interpolate the training data and the forecast data as 5° × 5°. For the 3D-Geoformer, these data are interpolated to regular grids with a resolution in the zonal direction of 2° and in the meridional direction of (1°) 0.5° (out of) 5° S to 5° N. The region of the data that DL models used is 0°–360°, 20° N–20° S. Note here we do not add the observation to the training data as the improvement of ENSO prediction skill is limited (Supplementary Fig. 10 and Supplementary Discussion 1 in ref. 31).

In order to compare the prediction skill of ENSO between the dynamical model with DL models, we use the historical ensemble hindcast data from eight fully coupled models in the NMME. The specific models are CanCM4i, CanSIPS-IC3, CanSIPSv2, COLA-RSMAS-CCSM3, COLA-RSMAS-CCSM4, GFDL-CM2p5-FLOR-A06, GFDL-CM2p5-FLOR-B01 and GFDL-CM2p5-aer04. The hindcast period is from 1982–2018. More details can be seen in the ref. 20. Note here we choose these eight models because they have the forecast data from 1982 to 2018, and the lead time is up to one year.

All monthly data are used in this study after the climatological seasonal cycle and linear trends have been removed.

Convolutional neural network model
The CNN model we used for training is the model developed by the ref. 29. and do not modify its architecture. Based on this architecture, we only use SST and SSH fields from CMIP6 as the training data. We then use this model to forecast the Niño3.4 index. The network architecture of the CNN model consists of one input layer, three convolutional layers, two pooling layers, one fully connected layer, and one output layer. The maximum pooling process extracts the maximum value from each 2×2 grid. The third convolutional layer is connected to the neurons of the fully connected layer, and the fully connected layer is connected to the final output. The input data are the SST and SSH fields for the previous three months with different lead months for the target month, and the output data are the scalar Niño3.4 index for the target month. The total number of convolutional filters and neurons in the fully connected layer is either 30 or 50. Thus, a combination of four CNN models can be obtained (C30H30, C30H50, C50H30, C50H50, where the numbers after C and H denote the number of convolutional filters (i.e., C) and neurons (i.e., H) in the fully connected layer, respectively). We take the average of the final four combinations as the result of the CNN model output, which also reduces the model's prediction error and makes the prediction more accurate. The size of the mini batch for each epoch is set to 400, and the number of epochs is 700 for the training using CMIP6 output.

3D-Geoformer model
The 3D-Geoformer model we used for training was developed by ref. 31. We do not modify the architecture of this model. This 3D-Geoformer model is built on an encoding-decoding strategy with associated modules, which includes two data preprocessing modules, encoding and decoding modules, and an output layer (more details in ref. 31). In contrast to ref. 31, which used wind fields, SST, and upper-ocean temperature anomalies fields (92°E-30°W, 20° S–20° N) as training data, we use only SST and SSH (0°–360°, 20° S–20° N) of CMIP6 for training, which are consistent with the training data in CNN. Although we only use SST and SSH data for training, the prediction skill of ENSO is similar to ref. 31. The 3D-Geoformer model takes 12 consecutive months of gridded SST and SSH fields as input data. The output data are gridded SST and SSH fields for the next 20 lead months.

Strategy 1
For the eight NMME dynamical models, we find that the prediction skill of the dynamic-model mean is higher than the prediction skill of any single model. Inspired by this, we get the final ENSO forecast results by averaging the dynamical multi-model, 3D-Geoformer, and CNN forecasts. Then, we calculate the ACC of these forecast results, which we designate as the ACC of Strategy 1.

Strategy 2
For a dynamical model, to avoid the perturbation of the initial field data on the forecast results of the model, several initial fields are used to form corresponding model members. We then average the simulation results of these members to get the final forecast results. However, this simple ensemble averaging method may pull down the prediction skill of the model[36,37]. Ref. 36 showed that the winter prediction skill of the North Atlantic Oscillation can be significantly improved by refining a dynamical ensemble through subsampling. They develop an approach called "First-Guess". Firstly, they propose a statistical approach and make the winter prediction based on the observed autumn fields as the "First-Guess" indicator. Then, they select 10 out of 30 members of the dynamical model, which are the closest to the "First-Guess" prediction. They use a subsampling approach from
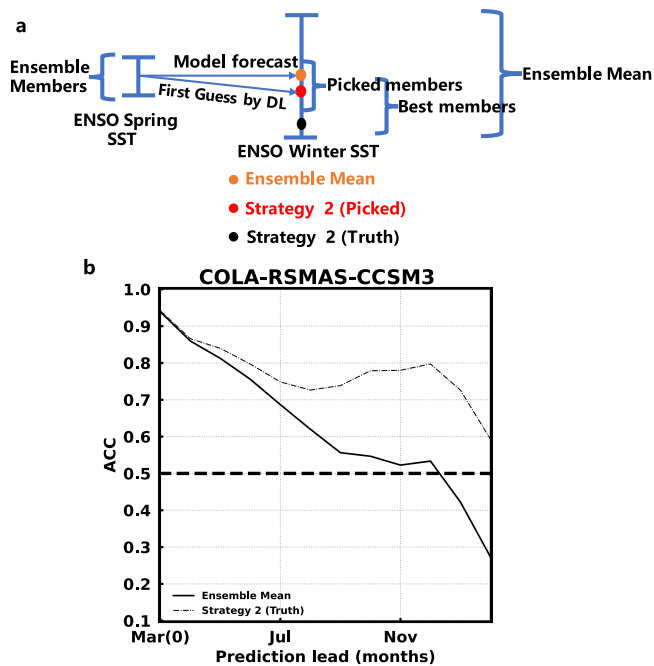
Fig. 4 | Strategy 2. a Schematic representation of Strategy 2 (see "Methods"). **b** One of the dynamical models, ACC, is used when the prediction is made from March during 1982–2018. The thin solid line indicates the ACC of the all-ensemble mean. The dashed-dot line indicates the ACC based on Strategy 2 using "Truth" value as an indicator (sub-ensemble members are 50% of all-ensemble members).

all-ensemble members and get the sub-ensemble members based on the "First-Guess" indicator. Similarly, ref. 37. Applied this method for the North Atlantic Oscillation study. These can be found that the forecast result of the sub-ensemble is significantly higher than that of all ensemble means. Inspired by this study, we use the DL models' prediction as the "First-Guess" indicator in ENSO prediction (Fig. 4a, take the forecast starting in March as an example). When the dynamical model forecasts winter ENSO with the initial month in the spring, the DL models also do the predictions with the same initial months and lead months. The Niño3.4 index obtained from the DL models forecasts for the months of December, January, and February (DJF) is used as the "First-Guess" indicator (noted as DL_{DJF}). Similarly, we calculate the DJF prediction of the Niño3.4 index for each member of the dynamical models made from spring (noted as $Model_{DJF}$). And then, we use this subsampling approach to select from all ensemble members to obtain the sub-ensemble members according to the smallest difference between DL_{DJF} and $Model_{DJF}$. Note that for each dynamical model, there is a different number of members, and we select 50% of all members of each dynamical model whose $Model_{DJF}$ is closest to the DL_{DJF} as the sub-ensemble member. We average the forecast results of all-ensemble members and sub-ensemble members and calculate their ACC, respectively (noted as Ensemble Mean and Strategy 2 (Picked), brown dot and red dot in Fig. 4a, respectively). Similarly, if we use the "Truth" (observations; black dot in Fig. 4a) value as a "First-Guess" indicator, we can get the Best 50% Mean according to the subsampling approach mentioned above. It is the upper predictability limit (i.e., an upper limit, in practice, cannot be achieved but can be asymptotically approached) of the prediction skill made by Strategy 2, as we cannot know the future in advance. This method can also be applied to the other seasons, such that we can calculate year-round ACC (i.e., Fig. 3).

We use the "First-Guess" for one of the dynamical models (Fig. 4b) as an example. Note that we only show the Ensemble Mean (solid line in Fig. 4b) and Strategy 2 (we use the corresponding months of the observation as the indicator). We can find that the ACC of Strategy 2

(Truth) is significantly higher than the Ensemble Mean. Therefore, there is still some room for using the "First-Guess" to improve the prediction skill.

Definition of ACC and RMSE
In this study, the prediction skill is quantified using the ACC. ACC is defined as the temporal anomaly correlation coefficient between the ensemble mean forecast (F_i) and the corresponding "Truth" (O_i, i.e., observation). RMSE is defined as the root mean square error between the ensemble mean forecast (F_i) and the corresponding "Truth" (O_i, i.e., observation).

$$ACC = \frac{<F_i, O_i>}{\sqrt{<F_i, F_i><O_i, O_i>}} \qquad (1)$$

$$RMSE = \sqrt{<F_i - O_i, F_i - O_i>} \qquad (2)$$

where F_i is the ensemble mean forecast anomaly for forecast month or year i, and O_i is the verifying observed anomaly. <> denotes the variance over all the months or years in verifying time series.

Bootstrap
The confidence interval of the forecast skills for the DL, dynamical models, Strategy 1, and Strategy 2 is calculated using the bootstrap method. At first, we randomly select N ensemble members, where N represents the number of ensemble members for each forecast system (e.g., N is 10 for the COLA-RSMAS-CCSM4 model; Supplementary Table 2). Overlapping is permitted during this random selection, meaning a selected ensemble member can be chosen more than once. The forecast skill of the ensemble-averaged value is then calculated. This procedure is repeated 1,000 times, and the 25th highest and lowest forecast skill values are used to define the 95% confidence interval.

Data availability
All data related to this paper can be downloaded as follows: The ORAS5 data are available at https://cds.climate.copernicus.eu/cdsapp# !/dataset/reanalysis-oras5?tab=form. The CMIP6 data can be downloaded online https://esgf-node.llnl.gov/projects/cmip6/. The SODA version 2.2.4, https://climatedataguide.ucar.edu/climate-data/soda-simple-ocean-data-assimilation. The NMME data are available at http://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/. Source data to reproduce the figures of this paper are available on https://doi.org/10.5281/zenodo.15162425.

Code availability
Code for the main results is available on https://doi.org/10.5281/zenodo.15162425.

References
1. McPhaden, M. J., Zebiak, S. E. & Glantz, M. H. ENSO as an integrating concept in earth science. *Science* **314**, 1740–1745 (2006).
2. Cai, W. et al. Increased variability of eastern Pacific El Niño under greenhouse warming. *Nature* **564**, 201–206 (2018).
3. Henson, C., Market, P., Lupo, A. & Guinan, P. ENSO and PDO-related climate variability impacts on midwestern United States crop yields. *Int. J. Biometeor.* **61**, 857–867 (2017).
4. Lehodey, P. et al. ENSO impact on marine fisheries and ecosystems. *El Niño South. Oscillation Changing Clim.* **19**, 429–451 (2020).
5. Liu, Y., Cai, W., Lin, X., Li, Z. & Zhang, Y. Nonlinear El Niño impacts on the global economy under climate change. *Nat. Commun.* **14**, 5887 (2023).
6. Yang, S. et al. El Niño-southern oscillation and its impact in the changing climate. *Natl. Sci. Rev.* **5**, 840–857 (2018).

7. Webster, P. J. & Yang, S. Monsoon and ENSO: selectively interactive systems. *Quart. J. Roy. Meteor. Soc.* **118**, 877–926 (1992).

8. Liu, Z., Jin, Y. & Rong, X. A theory for the seasonal predictability barrier: threshold, timing, and intensity. *J. Clim.* **32**, 423–443 (2019).

9. Jin, Y., Lu, Z. & Liu, Z. Controls of spring persistence barrier strength in different ENSO regimes and implications for 21st century changes. *Geophys. Res. Lett.* https://doi.org/10.1029/2020GL088010 (2020).

10. Jin, Y., Liu, Z. & McPhaden, M. J. A theory of the spring persistence barrier on ENSO. Part III: the role of tropical Pacific Ocean heat content. *J. Clim.* **34**, 8567–8577 (2021).

11. Jin, Y., Liu, Z. & Duan, W. The different relationships between the ENSO spring persistence barrier and predictability barrier. *J. Clim.* **35**, 6207–6218 (2022).

12. Luo, J.-J., Masson, S., Behera, S. & Yamagata, T. Extended ENSO predictions using a fully coupled ocean-atmosphere model. *J. Clim.* **21**, 84–93 (2008).

13. Barnston, A. G., Tippett, M. K., L'Heureux, M. L., Li, S. & DeWitt, D. G. Skill of real-time seasonal ENSO model predictions during 2002–11: Is our capability increasing? *Bull. Am. Meteor. Soc.* **93**, 631–651 (2012).

14. Ren, H. & Jin, F. F. Recharge oscillator mechanisms in two types of ENSO. *J. Clim.* **26**, 6506–6523 (2013).

15. Ren, H. et al. The new generation of ENSO prediction system in Beijing Climate Centre and its predictions for the 2014/2016 super El Niño event. *Meteor. Mon.* **42**, 521–531 (2016).

16. Zhang, R. H. & Gao, C. The IOCAS intermediate coupled model (IOCAS ICM) and its real-time predictions of the 2015–16 El Niño event. *Sci. Bull.* **66**, 1061–1070 (2016).

17. Tang, Y. et al. Progress in ENSO prediction and predictability study. *Nat. Sci. Rev.* **5**, 826–839 (2018).

18. Cane, M. A., Zebiak, S. E. & Dolan, S. C. Experimental forecasts of El Niño. *Nature* **321**, 827–832 (1986).

19. Zebiak, S. E. & Cane, M. A. A model El Niño-southern oscillation. *Mon. Wea. Rev.* **115**, 2262–2278 (1987).

20. Kirtman, B. P. et al. The North American multimodel ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bull. Am. Meteor. Soc.* **95**, 585–601 (2014).

21. Penland, C. & Magorian, T. Prediction of Niño-3 sea surface temperatures using linear inverse modeling. *J. Clim.* **6**, 1067–1076 (1993).

22. Penland, C. A stochastic model of IndoPacific sea surface temperature anomalies. *Phys. D.* **98**, 534–558 (1996).

23. Tseng, Y. H., Hu, Z. Z., Ding, R. & Chen, H. C. An ENSO prediction approach based on ocean conditions and ocean–atmosphere coupling. *Clim. Dyn.* **48**, 2025–2044 (2017).

24. Petrova, D., Ballester, J., Koopman, S. J. & Rodó, X. Multiyear statistical prediction of ENSO enhanced by the tropical Pacific observing system. *J. Clim.* **33**, 163–174 (2020).

25. Dijkstra, H. A., Petersik, P., Hernández-García, E. & López, C. The application of machine learning techniques to improve El Niño prediction skill. *Aip. Conf. Proc.* **7**, 478796 (2019).

26. Zhang, R. H., Gao, C. & Feng, L. Recent ENSO evolution and its real-time prediction challenges. *Nat. Sci. Rev.* **9**, 052 (2022).

27. Dong, C. et al. Recent developments in artificial intelligence in oceanography. *Ocean-Land-Atmos Res.* **2022**, 1–26 (2022).

28. Petersik, P. J. & Dijkstra, H. A. Probabilistic forecasting of El Niño using neural network models. *Geophys. Res. Lett.* **47**, e2019GL086423 (2020).

29. Ham, Y. G., Kim, J. H. & Luo, J. J. Deep learning for multi-year ENSO forecasts. *Nature* **573**, 568–572 (2019).

30. Hu, J. et al. Deep residual convolutional neural network combining dropout and transfer learning for ENSO forecasting. *Geophys. Res. Lett.* **48**, e2021GL093531 (2021).

31. Zhou, L. & Zhang, R. H. A self-attention-based neural network for three-dimensional multivariate modeling and its skillful ENSO predictions. *Sci. Adv.* **9**, 2827 (2023).

32. Lyu, P. M. et al. ResoNet: robust and explainable ENSO forecasts with hybrid convolution and transformer networks. *Adv. Atmos. Sci.* **41**, 1289–1298 (2024).

33. Zhao, S. et al. Explainable El Niño predictability from climate mode interactions. *Nature* **630**, 891–898 (2024).

34. Barnston, A. G., Tippett, M. K., Ranganathan, M. & L'Heureux, M. L. Deterministic skill of ENSO predictions from the North American multimodel ensemble. *Clim. Dyn.* **53**, 7215–7234 (2019).

35. Palmer, T. N. et al. Development of a European multimodel ensemble system for seasonal-to interannual prediction (DEMETER). *Bull. Am. Meteor. Soc.* **85**, 853–872 (2004).

36. Dobrynin, M. et al. Improved teleconnection-based dynamical seasonal predictions of boreal winter. *Geophys. Res. Lett.* **45**, 3605–3614 (2018).

37. Smith, D. M. et al. North Atlantic climate far more predictable than models imply. *Nature* **583**, 796–800 (2020).

38. Patil, K. R., Doi, T., Jayanthi, V. R. & Behera, S. Deep learning for skillful long-lead ENSO forecasts. *Front. Clim.* **4**, 1058677 (2023).

39. Bi, K. et al. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619**, 533–538 (2023).

40. Lam, R. et al. Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421 (2023).

Acknowledgements

Author contributions

Conceptualization: Y.J. and Z.L. Methodology: Y.J., Y.C. and X.S. Investigation: Y.C., Y.J., X.S., Z.L. and X.C. Visualization: Y.C. Funding acquisition: X.C. and Y.J. Project administration: Y.J., Z.L. and X.C. Supervision: Y.J., X.S., Z.L., X.C., X.L., R.-H.Z., J.L., W.Z., W.D., F.Z., M.M. and L.Z. Writing—original draft: Y.C., Y.J. and Z.L. Writing—review and editing: Y.J., Y.C., X.S., Z.L., X.C., R.-H.Z., W.Z., W.D., F.Z., M.M.

Competing interests

The authors declare no competing financial or non-financial interests.

Additional information