On the comparative utility of entropic learning versus deep learning for long-range ENSO prediction

Michael Groom[1], Davide Bassetti[2], Illia Horenko[3], and Terence J O'Kane[1]

[1]CSIRO Environment
[2]Faculty of Mathematics, Rheinland-Pfälzische, Technische Universität Kaiserslautern Landau
[3]Chair for Mathematics of AI, Rheinland-Pfälzische, Technische Universität Kaiserslautern Landau

April 16, 2024

On the comparative utility of entropic learning versus deep learning for long-range ENSO prediction

Michael Groom,[a] Davide Bassetti[b] Illia Horenko[c] and Terence J. O'Kane[d]

[a] *CSIRO Environment, Eveleigh, NSW 2015, Australia*

[b] *Faculty of Mathematics, Rheinland-Pfälzische Technische Universität Kaiserslautern Landau, Kaiserslautern 67663, Germany*

[c] *Chair for Mathematics of AI, Rheinland-Pfälzische Technische Universität Kaiserslautern Landau, Kaiserslautern 67663, Germany*

[d] *CSIRO Environment, Battery Point, TAS 7004, Australia*

Corresponding author: Michael Groom, michael.groom@csiro.au

ABSTRACT: This paper compares the ability of deep learning and entropic learning methods to predict the probability of the Niño3.4 index being above 0.4° (El Niño), below −0.4° (La Niña) or within both of these thresholds (neutral) at lead times of 3 up to 24 months. In particular, the performance, interpretability, and training cost of entropic learning methods, represented by the entropy-optimal Scalable Probabilistic Approximation (eSPA) algorithm, are compared with deep learning methods, represented by a Long Short-Term Memory (LSTM) classifier, trained on the same dataset. Using only data derived from observations over the period 1958-2018 and a corresponding surface-forced ocean model, the problem manifests as a canonical small-data challenge. Relative to the LSTM model, eSPA exhibits substantially better out-of-sample performance in terms of area under the ROC curve (AUC) for all lead times at $\sim 0.02\%$ of the computational cost. Comparisons of AUC with other state-of-the-art deep learning models in the literature show that eSPA appears to also be more accurate than these models across all three classes. Composite images are generated for each of the cluster centroids from each trained eSPA model at each lead time. At shorter lead times, the composite images for the most significant clusters correspond to patterns representing mature or emerging/declining El Niño or La Niña states, while at longer lead times they correspond to precursor states consisting of extra-tropical anomalies. Finally, modifications to the baseline dataset are explored, showing that improvements can be made in the parsimony of the trained eSPA model without sacrificing predictive power.

1. Introduction

a. Background and motivation

This paper is concerned with predicting variability in the climate system over seasonal to interannual time scales, specifically the El Niño-Southern Oscillation (ENSO). ENSO is characterised by irregularly periodic variations in sea surface temperature (SST) anomalies and trade winds over the tropical regions of the Pacific Ocean and is the dominant mode of interannual climate variability (Bjerknes 1969; Rasmusson and Carpenter 1982), typically taking between 3 and 7 years to transition from one mature phase to the other via the neutral phase (Neelin et al. 1998). ENSO variability, particularly the extremes of the positive (El Niño) and negative (La Niña) phases, has global impacts on climate, ecosystems and economies, making forecasts of ENSO particularly valuable for managing and mitigating these impacts. However, despite decades of research and development, forecasts of ENSO events at lead times longer than one year remain difficult to perform with any meaningful accuracy using conventional dynamical (i.e. physics-based) or statistical models (Barnston et al. 2012).

Recently, statistical approaches based on deep learning have shown some promise in producing skilful forecasts for lead times up to 18 months or longer (Ham et al. 2019), typically by leveraging large datasets such as the Coupled Model Intercomparison Project (CMIP) ensemble of climate projections using coupled ocean-atmosphere general circulation models (GCMs) (Taylor et al. 2012; Eyring et al. 2016). However, the use of historical simulations with unconstrained GCMs for learning ENSO variability is problematic given that the models comprising the CMIP ensemble exhibit large variations in the ENSO power spectrum, as well as in the causal interactions between different oscillatory components of the ENSO time series (Jajcay 2018). This paper presents an alternative approach based on recent developments in machine learning methods for small data problems (Horenko 2020; Vecchi et al. 2022), thereby circumventing the need to rely on big data and associated techniques such as transfer learning in order to produce models with meaningful skill at multi-year lead times. These methods, referred to as entropic machine learning, enable the sole use of observations or reanalyses that assimilate these observations for predicting the future evolution of the climate system. This is of great importance given the relatively short period of observations available, particularly for the ocean and its corresponding modes of variability, such

as ENSO. For example, since the beginning of the satellite era of ocean observations circa 1980, only three extreme El Niño events have occurred along with a handful of smaller amplitude events. This is the primary reason why training statistical models on observations alone has proven to be difficult.

b. Review of previous research

Methodologically, there are two main classes of models used for producing ENSO forecasts; dynamical models that simulate the coupled oceanic-atmospheric physics to varying degrees of fidelity and statistical models that aim to predict the evolution of one or more of the ENSO indices (e.g. the Niño3.4 index, the Southern Oscillation index etc). Barnston et al. (2012) provide a detailed assessment of the skill of both classes of models over the period of 2002-2011, showing that the dynamical models slightly (but statistically significantly) outperformed their statistical counterparts over that period. This was primarily due to the dynamical models producing more accurate forecasts when traversing the boreal spring predictability barrier (Jin et al. 2008), whereas forecasts whose lead times did not traverse the months of April to June were more equally successful among all models. Statistical models were also shown to suffer from slippage to a greater degree, which is the tendency for predicted transitions to lag observed transitions in the ENSO state due to a bias towards persistence. An up-to-date version of the ENSO prediction plume, featuring many of the same models of both classes as presented in the study by Barnston et al. (2012), can be found at the International Research Institute for Climate and Society (IRI) web page (`https://iri.columbia.edu/our-expertise/climate/forecasts/enso/current/?enso_tab=enso-sst_table`).

In recent years, there has been a resurgence of interest in the development of statistical forecast models for ENSO with the advent of deep learning. One of the earliest and most prominent examples is the study by Ham et al. (2019), which utilised transfer learning to train a convolutional neural network (CNN), first on historical simulations from the CMIP5 ensemble and then on data from the Simple Ocean Data Assimilation (SODA) reanalysis (Giese and Ray 2011). Comparisons with the SINTEX-F dynamical forecast system, as well as with various members of the North American Multi-Model Ensemble, showed that the forecast skill of the CNN model was superior to that of any of the dynamical forecast systems at lead times longer than 6 months. Furthermore, the

all-season correlation skill for the Niño3.4 index in the CNN model was above 0.5 for lead times of up to 17 months. A subsequent study by Ham et al. (2021) improved on this result by utilising a multitask learning framework, where the CNN model was extended to simultaneously predict the observed calendar month of the input, thus allowing a single model to be trained for all seasons and lead times. This led to an overall increase in skill, in particular for forecasts initiated in the boreal spring. Kim et al. (2022) also performed multitask learning for predicting both the Niño3.4 index and the observed calendar month, but with a different architecture consisting of three modules. The first module employed 3D receptive field blocks (2D + time) with convolution filters and residual connections to encode spatio-temporal patterns in the input data, the second module consisted of a stateful Long Short-Term Memory (LSTM) network with a spatial attention mechanism to learn the temporal order of long-term sequences from the encoding module and predict the Niño3.4 index for the next 23 months, while the final module was a classification module with two fully connected layers for predicting the observed calendar month of the input. When compared to the Ham et al. (2021) model, the correlation coefficient at 12 months lead time was improved by 5.8% and the prediction of the calendar month was improved by 13%. The authors also noted that an overall lack of training data is likely a barrier to further performance improvements since conventional data augmentation methods such as flipping, rotation and translation cannot be used for spatio-temporal climate data.

Other recent studies using the CMIP5 and/or CMIP6 ensembles to train deep learning methods for ENSO prediction include Zhou and Zhang (2023); Gao et al. (2023), which employed a Transformer-based architecture with self-attention rather than the convolutional and recurrent neural network architectures featured in earlier studies, along with other improvements. Rather than directly predict an index, the model used in these studies predicts 2D sea surface wind stress and 3D upper-ocean temperature anomaly fields (to 150m depth), using these same fields from the previous 12 months as inputs. Forecasts were found to be skilful up to 18 months lead time for the Niño3.4 index, however, the authors also noted that biases in the simulation data used for training resulted in regions of lower skill, such as the equatorial western Pacific. Transfer learning using reanalysis data was not performed in these studies, due to difficulties in applying the technique for the high-dimensional inputs being used. Qiao et al. (2023) used a deep residual network with spatial attention in each residual block to predict the Niño3.4 index up to 24 months lead time. Inputs

to the network were the previous 12 months of SST anomalies as well as first- and second-order differences of these anomalies to provide additional temporal features for learning, referred to as tendency. Similar performance was obtained to previous studies in terms of correlation skill, with improved forecasts of extreme events. Wang et al. (2023) used a convolutional LSTM (ConvLSTM) with self-attention to predict the Niño3.4 index, with skilful forecasts obtained out to lead times of 20 months. This study utilised a genetic algorithm to filter the CMIP dataset prior to training by choosing the optimal combination of CMIP models to use from the entire ensemble. This strategy, along with the self-attention module in the ConvLSTM, was responsible for most of the performance improvements. Most recently, Wang and Huang (2024) used SST anomalies, as well as tendency features similar to Qiao et al. (2023), as inputs to a CNN model trained to predict the principal components (PCs) of the first three empirical orthogonal function (EOF) modes of SST anomalies in the tropical Pacific. Combining the predicted PCs with their corresponding EOFs allowed the authors to examine detailed maps of the precursors used to predict specific events.

An example of a method that bridges the gap between training on simulation output and training on observations is the study by Chen et al. (2021), who proposed a method that uses simulation output from an approximate parametric model, in this case the recharge-discharge model of Jin (1997) augmented with a random wind burst model, as the prior information while the observational data plays the role of the likelihood which corrects the intrinsic model error in the prior data during training of a feedforward neural network. The method makes use of two loss functions; the first involves the error between the network outputs and the simulation data, while the second involves the error between the network outputs and the observational data. The first loss function is used at each step of the gradient descent optimisation to propose an update to the network parameters. This proposal is then used to evaluate the second loss function and is rejected if it does not produce a decrease in this validation loss. The authors applied this method to predicting the Niño3 index and obtained skilful forecasts up to 10 months lead time. They also noted the absence of any spring predictability barrier, with forecast skill remaining at 10 months when initiated from any time between February and August.

There are also several recent studies worth mentioning here that solely use observational or reanalysis data for training various deep learning methods. Taylor and Feng (2022) trained a UNet-LSTM, consisting of ConvLSTM modules in an encoder-decoder architecture with skip

connections, to predict monthly mean sea surface temperature and 2m air temperature at lead times up to 24 months, using these same fields as inputs. Training data was taken from the ERA5 reanalysis over the period of 1950–2021. The authors found that while their model was skilful in predicting the 2019–2020 El Niño and the 2016–2017 and 2017–2018 La Niñas, it failed to predict the peak of the 2015–2016 El Niño, possibly due to the absence of any information about the subsurface ocean in their model. Chen et al. (2023) combined a seasonal-trend decomposition using locally weighted scatter plot smoothing to the Niño3.4 index, derived from the HadISST dataset over the period of 1871-2022, with temporal convolutional networks to perform multi-step predictions of each component in the decomposition (trend, seasonal and remainder), which were then combined to produce the final forecast for the index. Skilful forecasts were obtained out to 14 months lead time, while a similar model using the same decomposition but with LSTM modules in place of the temporal convolutional networks was able to achieve skilful forecasts out to 12 months. Finally, Patil et al. (2023) used a CNN model with heterogeneous parameters for each season, as well as a modified loss function that contained an extra penalty for failing to correctly predict extreme events. In contrast to the models used in Ham et al. (2019, 2021), each convolutional layer in the model was followed by dropout, regularisation and batch-normalisation layers as well as an average pooling layer to reduce the number of model parameters. Training data consisted of SST anomalies taken from the Centennial in situ Observation-Based Estimates dataset as well as vertically averaged subsurface temperature anomalies (averaged over 0–300m depth) taken from SODA, with the NOAA Optimum Interpolation SST and the NCEP Global Ocean Data Assimilation System datasets used for validation. Dimension reduction was performed by re-gridding the data to $5^\circ \times 5^\circ$ resolution. Despite the much smaller training dataset used, skilful forecasts were obtained out to 20 months lead time, compared with just 12 months lead time for SINTEX-F2 and a fixed parameter CNN model. The authors also evaluated the probabilistic skill of their model at 18 and 23 months lead time using the area under the ROC curve (AUC) of true vs. false positive predictions of each phase of ENSO. AUCs of 0.75, 0.75 and 0.62 were obtained for El Niño, La Niña and neutral phases at 18 months lead time, with AUCs of 0.69, 0.7 and 0.64 obtained at 23 months lead time respectively.

Despite the success of deep learning when applied to ENSO prediction, the limitations of climate model biases leading to biases in the training data and/or lack of sufficient observations motivate

the search for other algorithms that may be able to overcome these limitations. A promising alternative class of machine learning methods, developed specifically to avoid overfitting for small data problems (i.e. problems where the number of features is of similar size or even greater than the number of data instances available for training), has been proposed in Horenko (2020); Vecchi et al. (2022); Horenko et al. (2023). This study will focus specifically on applying the entropy-optimal Scalable Probabilistic Approximation (eSPA) classifier first presented in Horenko (2020) and then further improved in Vecchi et al. (2022). In addition to the various algorithmic improvements presented in Vecchi et al. (2022), as well as favourable comparisons with other ML methods on synthetic data, this study was also the first to apply eSPA to ENSO prediction by formulating the problem as a classification task. Specifically, eSPA was employed to predict whether the Niño3.4 index was above or below the threshold value of 0.4° (used to define the presence of an El Niño event) at a lead time of 24 months. The training data consisted of the observed index as well as the first 100 PCs from an EOF analysis of global SST anomalies, along with the first 100 PCs from an EOF analysis of the vertical derivative of meridionally averaged water temperature at the equator to a depth of 500m, taken from a resimulated ocean model dataset (O'Kane et al. 2014). Substantially better performance was obtained compared to a benchmark LSTM model trained on the same data, with eSPA predicting instances on the test set with 87% accuracy (and an AUC of 0.82) compared to only 61% accuracy for the LSTM model (AUC of 0.49). In a subsequent study, Horenko et al. (2023) formulated the Sparse Probabilistic Approximation for Regression Task Analysis (SPARTAn) algorithm in order to directly predict the Niño3.4 index and compared it with various other methods, including an LSTM model. Substantially lower MSE values were obtained by SPARTAn than the other methods out to lead times of 15 months, however forecast skill (as measured by the pattern correlation for the test set) was not presented. Based on these findings, the current paper aims to perform a much more detailed study to assess the abilities of eSPA for predicting ENSO variability, including the various measures of interpretability that are enabled by the formulation of the algorithm. A follow-up paper will also perform a similar analysis for SPARTAn.

c. *Outline of paper*

The remainder of this paper is organised as follows. Section 2 describes the eSPA algorithm in detail, along with the dataset used for training. Section 3 presents out-of-sample predictions for lead times of 3, 6, 12, 18 and 24 months, using both eSPA and a benchmark LSTM method. Section 4 explores various modifications to the baseline results presented in Section 3, showing how further improvements can be made to both the performance and parsimony of eSPA for the ENSO prediction task. Finally, Section 5 contains conclusions and a discussion of directions for future work.

2. Methodology and dataset preparation

The data-driven predictions made throughout this paper make use of the recently proposed eSPA+ algorithm (Vecchi et al. 2022), with an LSTM classifier used as a benchmark trained on the same dataset. Here, prediction of the El Niño-Southern Oscillation is formulated as a classification task, following the example presented in Vecchi et al. (2022). However, unlike Vecchi et al. (2022) the dataset is extended to contain multiple classes; the Niño3.4 index is coarse-grained to take a value of 1 if it exceeds +0.4°, -1 if it exceeds −0.4° and 0 otherwise (labelled as classes 3, 2 and 1 respectively). This is a more challenging prediction task than the binary classification formulation used in Vecchi et al. (2022), but one which enables the prediction of both El Niño and La Niña events. For lead times ranging from 3 months to 24 months, a classifier is trained to predict the labelled data based on a set of features derived from a resimulated ocean dataset over the period of 1958 to 2018.

The training data consists of a feature matrix $X \in \mathbb{R}^{D\times T}$ and a label matrix $\Pi \in \mathbb{R}^{M\times T}$, where D is the number of features, T is the number of data instances and M is the number of labels. $\Pi_{m,t}$ represents the probability that $X_{:,t}$ belongs to class $m \in [1, M] \cap \mathbb{Z}$, taken here to be a hard label (i.e. each instance can only belong to a single class with probability 1). This classification task represents an example of supervised machine learning in the small data regime, since the number of features D is of similar size to the number of data instances T available for training (given that the ocean data is provided in monthly intervals). The recently proposed eSPA classifier has been shown to avoid overfitting in this regime (Horenko 2020; Vecchi et al. 2022). eSPA simultaneously

performs clustering, feature selection and classification and, as will be demonstrated in this paper, is physically interpretable.

a. The eSPA+ algorithm

The entropy-optimal Scalable Probabilistic Approximation algorithm is based on the Scalable Probabilistic Approximation (SPA) algorithm introduced by Gerber et al. (2020) for unsupervised learning problems. SPA aims to find an optimal discretisation of the feature matrix X by introducing a segmentation matrix $S \in \mathbb{R}^{D\times K}$, with $S_{:,k}$ being the centroid of cluster $k \in [1, K] \cap \mathbb{Z}$, along with an affiliation matrix $\Gamma \in \mathbb{R}^{K\times T}$, where $\Gamma_{k,t}$ is the probability that $X_{:,t}$ is in cluster k. $\hat{X} = S\Gamma$ is referred to as the reconstruction of X, and the optimal discretisation is sought by minimising the following regularised functional,

$$\mathcal{L}_{\text{SPA}} = \frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T}\left(X_{d,t} - \{S\Gamma\}_{d,t}\right)^2 + \varepsilon_S \sum_{d=1}^{D}\sum_{k_1=1}^{K}\sum_{k_2=1}^{K}\left(S_{d,k_1} - S_{d,k_2}\right)^2, \tag{1}$$

subject to the constraints $\Gamma_{k,t} \in [0, 1]$ and $\sum_{k=1}^{K}\Gamma_{k,t} = 1\forall t$. The second term of the $\mathcal{L}_{\text{SPA}}$ functional, whose relative importance is regulated by the hyperparameter $\varepsilon_S \geq 0$, is a regularisation term that is included so as to minimise the distance between each of the cluster centroids, thus filtering out those dimensions that do not significantly impact the discretisation error (given by the first term). The K clusters, also called discretisation boxes, are piece-wise linear and disjoint and are chosen in such a way that they provide a tessellation of the feature space. In Gerber et al. (2020) it was proven that the optimal segmentation is strictly piece-wise linear, a result which holds for both a discrete segmentation ($\Gamma_{k,t} \in \{0, 1\}\forall k, t$) and a fuzzy segmentation ($\Gamma_{k,t} \in [0, 1]\forall k, t$), meaning that each instance can belong to multiple boxes with different probabilities. Furthermore, the constrained minimisation of equation 1 can also be achieved with a computational cost that scales linearly with D and T (Gerber et al. 2020).

The extension of SPA to supervised learning problems was introduced in Horenko (2020). The entropy-optimal SPA (eSPA) algorithm replaces the regularisation term in the $\mathcal{L}_{\text{SPA}}$ functional with an entropy-based filtering of the feature space and introduces an additional term representing the classification error on the labelled data (measured using the Kullback-Leibler divergence between the true class probabilities and the predicted class probabilities). The regularised loss functional

that is minimised in eSPA is given by

$$\mathcal{L}_{\text{eSPA}} = \frac{1}{T}\sum_{d=1}^{D} W_d \sum_{t=1}^{T} \left(X_{d,t} - \{S\Gamma\}_{d,t}\right)^2 + \varepsilon_E \sum_{d=1}^{D} W_d \log(W_d) - \frac{\varepsilon_C}{T}\sum_{m=1}^{M}\sum_{t=1}^{T} \Pi_{m,t} \log\left(\sum_{k=1}^{K} \Lambda_{m,k}\Gamma_{k,t}\right). \tag{2}$$

subject to the constraints $\Gamma_{k,t} \in [0,1]$, $\sum_{k=1}^{K}\Gamma_{k,t} = 1 \forall t$, $W_d \in [0,1]$, $\sum_{d=1}^{D} W_d = 1$, $\Lambda_{m,k} \in [0,1]$ and $\sum_{m=1}^{M}\Lambda_{m,k} = 1 \forall k$. Compared to SPA, the average discretisation error over all features $d = 1, \ldots, D$ is now weighted by a vector $W \in \mathbb{R}^D$, where W_d represents the probability that feature d contributes to the discretisation error. The first regularisation term, controlled by the hyperparameter $\varepsilon_E \geq 0$, maximises the entropy of W to give the least biased estimate (subject to all other constraints), in accordance with the principle of maximum entropy (Jaynes 1957a,b). In the limit $\varepsilon_E \to \infty$, W converges to the uniform distribution with $W_d = 1/D$ and the previous SPA discretisation is obtained. The second regularisation term, controlled by the hyperparameter $\varepsilon_C \geq 0$, minimises the classification error that is obtained when representing the relationship between discretisation boxes and labels as a Bayesian network. This relationship is expressed via the law of total probability as $\hat{\Pi} = \Lambda\Gamma$, with the matrix $\Lambda \in \mathbb{R}^{M\times K}$ containing the conditional probabilities $\Lambda_{m,k}$ that $X_{:,t}$ belongs to class m, conditional on being in cluster k. The second regularisation term is then the cross-entropy loss between the true class probabilities Π and the predicted class probabilities $\hat{\Pi}$ since minimising this is equivalent to minimising the Kullback-Leibler divergence in the case where the true probabilities are constant. Note that setting $\varepsilon_C = 0$ results in a solution of the unsupervised discretisation and feature selection problems only.

Theorem 1 in Horenko (2020) summarises the monotonicity of convergence to, and regularity of, the optimal solution, as well as the computational complexity of the iterative numerical algorithm used to minimise the loss functional. For each iteration, the eSPA algorithm consists of four consecutive substeps, each obtained by solving a convex optimisation problem with three of the four unknowns fixed and each of which monotonically decreases the value of the loss functional given by equation 2. In its original formulation, the algorithm does not scale linearly with D and T since the substep pertaining to the calculation of the probability vector W relied on an interior point method with complexity $\mathcal{O}(D\log(D) + KT[M+D])$, regardless of whether the affiliation matrix Γ is discrete or fuzzy. In Vecchi et al. (2022), an improved algorithm eSPA+ was proposed involving a reordering of the optimisation substeps along with the derivation of closed-form solutions to

each of the substeps for the case of a discrete segmentation (i.e. $\Gamma_{k,t} \in \{0,1\} \forall k,t$). In this case, by deploying Jensen's inequality (Jensen 1906), the eSPA+ loss functional can be rewritten as

$$\mathcal{L}_{\text{eSPA+}} = \frac{1}{T}\sum_{d=1}^{D} W_d \sum_{t=1}^{T} \Gamma_{k,t}\left(X_{d,t} - S_{d,k}\right)^2 + \varepsilon_E \sum_{d=1}^{D} W_d \log(W_d) - \frac{\varepsilon_C}{T}\sum_{m=1}^{M}\sum_{t=1}^{T}\Pi_{m,t}\sum_{k=1}^{K}\Gamma_{k,t}\log(\Lambda_{m,k}), \tag{3}$$

subject to the same constraints as equation 2. Note that although the closed-form solution for the W substep does not depend on whether the segmentation is discrete or fuzzy, closed-form solutions for the Γ, S and Λ substeps may only be obtained for a discrete segmentation (see Horenko (2020); Vecchi et al. (2022) for further details). Furthermore, due to Jensen's inequality, the $\mathcal{L}_{\text{eSPA+}}$ loss functional is an upper bound on the $\mathcal{L}_{\text{eSPA}}$ loss functional. Therefore, even if the optimal Γ that minimises $\mathcal{L}_{\text{eSPA}}$ is fuzzy, minimisation of $\mathcal{L}_{\text{eSPA+}}$ will still provide an approximate solution. Multiple random restarts are used to help avoid getting trapped in a local minimum that does not provide good generalisation to unseen data. An additional improvement to the algorithm, presented here for the first time, involves discarding any empty boxes k after the calculation of each Γ substep. This improves the speed of the algorithm and decreases the number of iterations required for convergence since such boxes will always remain empty if they are empty initially due to the random initial choice of W, S and Λ. For brevity, the eSPA+ algorithm will simply be referred to as eSPA throughout the remainder of this paper.

b. Long Short-Term Memory Classifier

For comparison with eSPA, a Long Short-Term Memory classifier is trained to provide a benchmark representative of state-of-the-art deep learning methods *for the same dataset*, i.e. the same features as eSPA (described below) are used for prediction along with the same quantile transformation pre-processing step. Following Vecchi et al. (2022), the LSTM model architecture consists of D sequence input layers, a choice of 2, 4, 8, 16, 32, 64, 128 or 256 hidden units in the LSTM layer (determined using a grid search) and 3 hidden units in the final fully connected layer (one for each class), which uses the softmax activation function. The model is trained using the ADAM optimisation algorithm for 100,000 epochs with an initial learning rate of $\eta = 0.001$, a learning rate schedule that decreases η by 5% every 1000 epochs and the cross-entropy loss. Regularisation of

the model is performed using l^2 weight decay with a choice of the regularisation constant λ equal to 0, 1×10^{-6}, 1×10^{-5}, 1×10^{-4}, 1×10^{-3}, or 1×10^{-2} (also determined using a grid search).

c. Description of the dataset

The features used for prediction are the first 100 PCs from an EOF analysis of global SST anomalies, along with the first 100 PCs from an EOF analysis of the vertical derivative of water temperature (dT/dz) at the equator, averaged over latitudes of ±5°, as a proxy for thermocline variability. A pre-processing step of mapping the data to a uniform distribution with values between 0 and 1 using a quantile transformation is applied to all of the features.

The EOF analysis is performed for the ACCESS-OM2 resimulated ocean model dataset (Kiss et al. 2020), which uses JRA55-do interannual forcing over a period of 1958-2018. This is an updated version of the ACCESS-O model that was used in the previous studies on ENSO prediction using eSPA (Vecchi et al. 2022) and SPARTAn (Horenko et al. 2023). For comparison, both models have a nominal grid resolution of $1^\circ\times1^\circ$, with refinement to $1/3^\circ$ for latitudes between ±10°, a tripolar Arctic north of 65°N and a Mercator (cosine dependent) implementation for the Southern Hemisphere, ranging from $1/4^\circ$ at 78°S to 1° at 30°S. In the vertical direction, ACCESS-O has 50 levels covering 0–6000 meters with a grid spacing ranging from 10 meters in the upper layers (0–200 meters) to 333 meters for the abyssal ocean, whereas ACCESS-OM2 has 50 levels with 2.3m spacing at the surface, increasing smoothly to 219.6m spacing at the maximum depth of 5363.5m. Note that although there is no assimilation of subsurface ocean data (due to a lack of high-quality data prior to the ARGO era circa 2004), the model's representation of the subsurface ocean is still expected to be constrained by the forcing (and therefore by observations) in the upper portions of the ocean in the tropics, as was observed for the earlier ACCESS-O model (O'Kane et al. 2014).

A reference period of 1958-2012 (i.e. the first 90% of the dataset used for training) is used to calculate the monthly climatological means required for calculating the SST anomalies. The first 10 EOFs and PCs for SST and dT/dz are given in the supplementary material. Composite images of SST and dT/dz, generated by averaging over all instances for which the model Niño3.4 index was $> 0.4^\circ$, $< -0.4^\circ$ and $[-0.4^\circ, 0.4^\circ]$ to produce a canonical El Niño, La Niña and neutral phase composite respectively, are shown in the supplementary material.

The targets for prediction (labels) are generated by considering the probability of the Niño3.4 index exceeding 0.4°C (El Niño) or −0.4°C (La Niña) in n months time. The Niño3.4 index is produced using observed SST anomalies from the Hadley Centre Sea Ice and Sea Surface Temperature (HadISST) dataset (Rayner et al. 2003), with the same common reference period of 1958-2012 used to calculate the monthly climatological means. This is another point of difference with the previous results presented in Vecchi et al. (2022), which used a Niño3.4 index generated from the ACCESS-O model rather than observations. A comparison with the Niño3.4 index calculated from the ACCESS-OM2 model is shown in the supplementary material, along with a plot of the 0-month ahead labels.

The baseline dataset, presented in section 3, consists of features generated through principal component analysis (PCA) of the ACCESS-OM2 model output along with labels generated from the HadISST observational product as described above. In section 4, various sensitivity studies are conducted whereby either the features or labels are modified to see what effect this has on the clustering and predictions made by eSPA. In section 4a Singular Spectrum Analysis (SSA) with embedding dimensions of 3 and 6 months is used to generate the features rather than PCA. In section 4b two modifications to the labels are considered; (i) a doubling of the threshold to ±0.8°C and (ii) filtering out events that do not persist for at least 5 months. Finally, in section 4c the sensitivity to the amount of training data is explored by reducing the dataset in 10% increments until only 50% of the original training data remains. The data is removed from the beginning of the dataset so that the start date when only 50% of the original data remains is January 1988.

d. Performance metrics

For each lead time, classifier performance is assessed through the (macro-averaged) area under the ROC curve (AUC), the mean accuracy of predictions (ACC) and the expected calibration error (ECE), defined as

$$\mathrm{ECE} = \sum_{n=1}^{N} \frac{|B_n|}{T} |\mathrm{acc}(B_n) - \mathrm{conf}(B_n)|, \tag{4}$$

where the predicted probabilities are divided into N evenly spaced bins B_n of size $|B_n|$ and $\mathrm{acc}(B_n)$ and $\mathrm{conf}(B_n)$ are the accuracy and confidence for each bin, defined as

$$\mathrm{acc}(B_n) = \frac{1}{|B_n|} \sum_{i \in B_n} \mathbb{1}(\hat{y}_i = y_i), \qquad \mathrm{conf}(B_n) = \frac{1}{|B_n|} \sum_{i \in B_n} \max(\hat{\Pi}_{:,i}), \tag{5}$$

with $\hat{y}_i$ and y_i represent the predicted and true label for instance i and $\mathbb{1}$ is an indicator function that evaluates to 1 if true and 0 otherwise. Similar to AUC, for multi-class problems an ECE is computed for each class in a one vs. the rest approach and a macro-averaged ECE is produced as the final metric.

Additional measures specific to eSPA that are considered are the number of features $\tilde{D}$ with probability (given by the W vector) greater than the maximum entropy limit of $1/D$, as well as the number of clusters $\tilde{K}$ that are well-supported by the data, meaning they can be associated with a significant p-value. For this latter measure, the (two-tailed) p-value is calculated by forming a contingency table between $\Gamma_{k,:}$ and $\Pi_{m,:}$ and using Fisher's exact test to calculate the probability of observing this particular arrangement of the data under the null hypothesis that either value of the label m (i.e. 0 or 1) is likely to be present in the instances assigned to cluster k. The p-value that is returned for each cluster is the one corresponding to the label with the highest conditional probability (given by $\mathrm{argmax}(\Lambda_{:,k})$) and the standard rejection threshold of 0.05 is used to determine significance. Both $\tilde{D}$ and $\tilde{K}$ (as well as the ratios $\tilde{D}/D$ and $\tilde{K}/K$ and the total number of clusters K) are considered to be informative as they are measures of parsimony which can be used for model selection, all else being equal. Another useful measure to consider is the number of switches between clusters, denoted here by C, as this is informative of how persistent the clusters are in time.

3. Baseline ACCESS-OM2 results

In this section, the out-of-sample performance of eSPA is assessed through a simple forecasting task; the first 90% of the dataset is used to train a model, which is then used to make predictions for the final 10% of the dataset. The features used for prediction are those obtained through PCA of the ACCESS-OM2 dataset. Model selection is performed through a grid search over hyperparameters, where for each hyperparameter combination K-fold cross-validation is used, with 10 folds used and

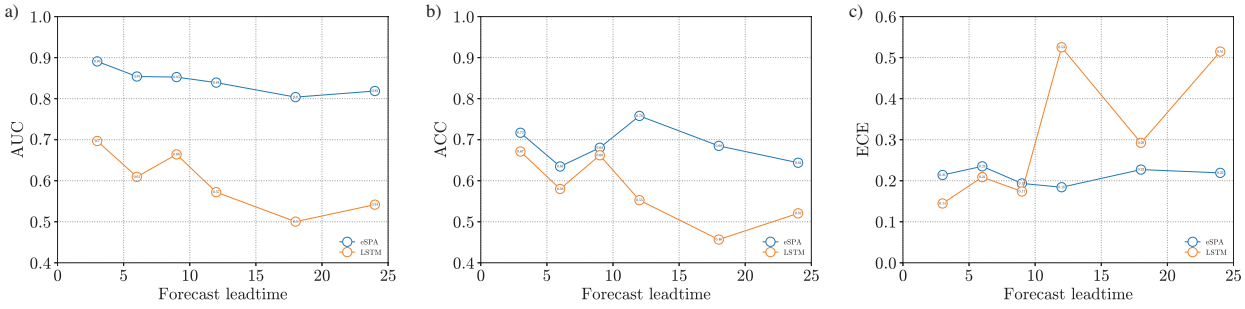


FIG. 1: Performance vs. lead time for the baseline dataset.

no permutation of the data prior to splitting. This increases variance in the predictions relative to other cross-validation strategies, such as repeated random subsampling, but reduces bias and helps to avoid overfitting by favouring models with fewer clusters. Only the first 9 folds (i.e. the training set) are used for the grid search, with the final fold reserved as the test set. The hyperparameters corresponding to the highest mean AUC from the grid search are then used to train the final model on the training set and make predictions for the test set.

a. Performance

Figure 1 shows plots of AUC, ACC and ECE vs. lead time for both eSPA and the benchmark LSTM method using the baseline dataset. At all lead times, the test set AUC is at least 0.8 for eSPA with a general trend of higher AUCs at shorter lead times. The mean accuracy of eSPA for each lead time is between 63% and 76% while the expected calibration error is between 0.18 and 0.24, with no noticeable trend observed with lead time. Note the AUC is the preferred metric over mean accuracy for scoring model performance on this problem due to both the imbalance of classes and the greater importance placed on correctly predicting departures from the neutral class. Indeed, based on the class priors, a model that just predicts the neutral class for all instances would score around 40% in terms of mean accuracy.

In comparison with eSPA, the LSTM model has a lower AUC and ACC at all lead times. For lead times of 3 to 9 months the LSTM mode performance is comparable to eSPA in terms of ACC and is better in terms of ECE, however for longer lead times the performance rapidly degrades and the gap between the two models increases.

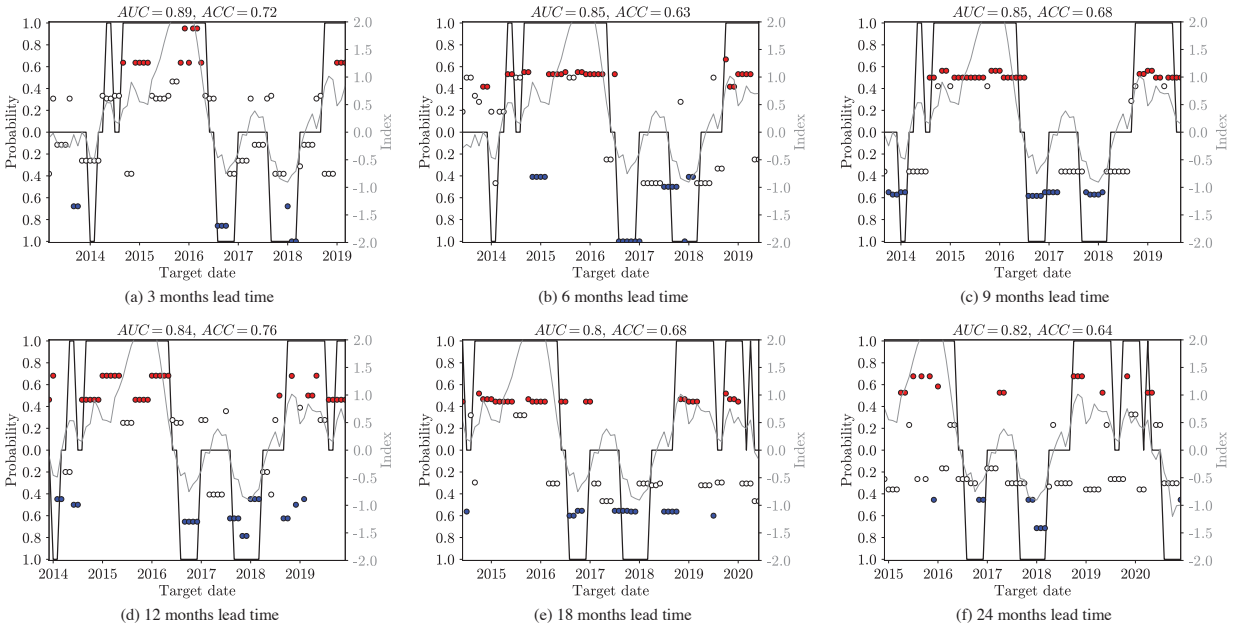
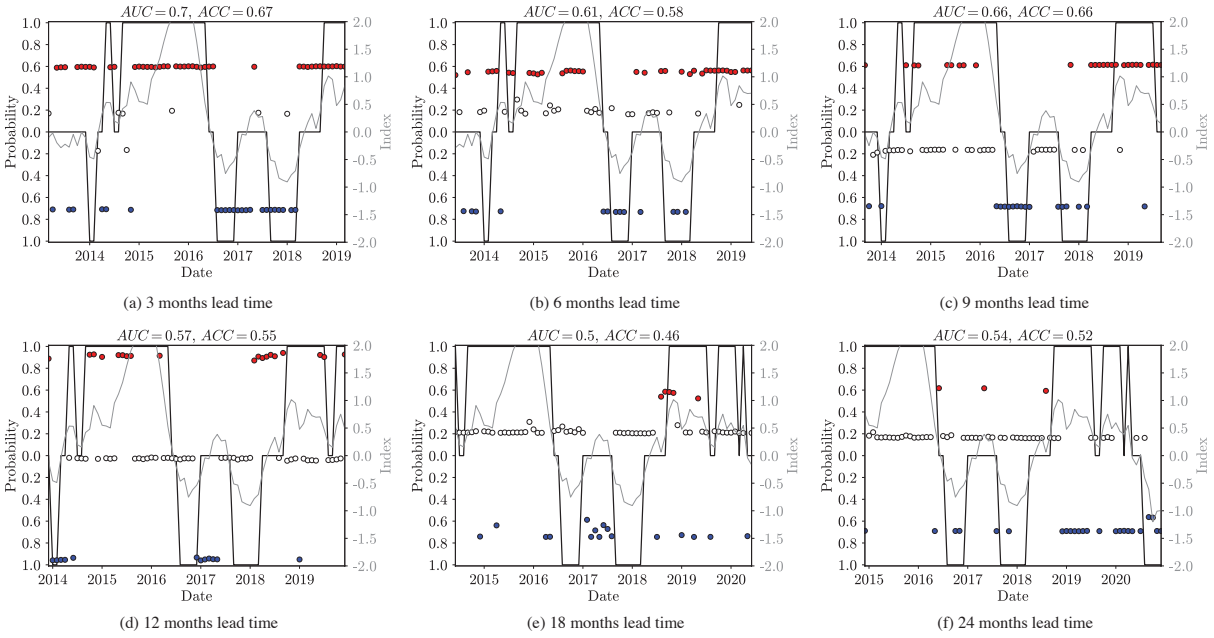


FIG. 2: eSPA test set predictions for the baseline dataset. Red circles indicate El Niño events, blue circles indicate La Niña events and white circles indicate neutral events. The Niño3.4 index (grey lines) and corresponding labels (black lines) for the target date are also shown.

Figure 2 shows the eSPA predictions on the test set for each lead time. The LHS y-axis for each plot shows the probability obtained from argmax $\left(\hat{\Pi}_{:,t}\right)$; in the case where this corresponds to an El Niño event (red circles) the probability is plotted positive upwards and for a La Niña (blue circles) it is plotted positive downwards. For neutral events (white circles), the second-highest probability is plotted using the same convention. The RHS y-axis shows the Niño3.4 index on the target date, calculated from HadISST observations. At all lead times, eSPA successfully predicts most of the high amplitude 2015/16 El Niño, the subsequent low amplitude La Niña events in 2016 and 2017/18 and the 2018/19 El Niño. By comparing to the Niño3.4 index at each target date, it is apparent that the value of the predicted probabilities of El Niño / La Niña are at least marginally correlated with the amplitude of an event. It is also encouraging to note that there are also only a handful of misclassified instances where an El Niño is classified as a La Niña and vice versa; in most cases the misclassifications are between the neutral state and El Niño or neutral and La Niña.



(a) 3 months lead time (b) 6 months lead time (c) 9 months lead time

(d) 12 months lead time (e) 18 months lead time (f) 24 months lead time

FIG. 3: LSTM test set predictions for the baseline dataset. Red circles indicate El Niño events, blue circles indicate La Niña events and white circles indicate neutral events. The Niño3.4 index (grey lines) and corresponding labels (black lines) for the target date are also shown.

b. Comparisons with benchmark

Figure 3 shows the equivalent LSTM predictions on the test set for each lead time. Note that unlike eSPA, the grid search over LSTM hyperparameters (in this case the number of hidden units and the l^2 regularisation constant) is performed using the AUC directly on the test set to score each candidate model. Therefore, the results shown here should be considered an optimistic estimate of the out-of-sample performance of LSTM on this dataset. The number of hidden units and regularisation constant in the final model are $(32, 0.1)$, $(8, 0.1)$, $(16, 0.1)$, $(2, 0.01)$, $(2, 0.01)$ and $(8, 0.1)$ for each lead time respectively. The corresponding number of parameters in each model is 29987 for 32 hidden units, 13971 for 16 hidden units, 6731 for 8 hidden units and 1637 for 2 hidden units.

As a general trend, at shorter lead times the model struggles to correctly predict the neutral phase, while at longer lead times the model struggles to predict phases other than neutral. Also, unlike eSPA, there is little variation in the probability assigned to each class, i.e. all predictions of El Niño have approximately the same probability of occurring and similarly for predictions of the

La Niña and neutral classes. In terms of computational cost, training the final LSTM model for 100000 epochs on a single CPU core took on average 135 seconds for models with 2 hidden units, 340 seconds for models with 8 hidden units, 420 seconds for models with 16 hidden units and 900 seconds for models with 32 hidden units. By comparison, training 10000 eSPA models (i.e. 10000 random restarts) on the same dataset took 300 seconds on average. In other words, an eSPA model can be completely trained (i.e. converged to a local minimum) for every 30–220 epochs of training the most competitive LSTM model on the same dataset. Both the LSTM model and eSPA were implemented entirely in the Julia programming language and were executed on a single core of an AMD EPYC 7543 32-core processor running at 2.8 GHz (3.7 GHz turbo) with 256 MB cache.

As mentioned in the introduction, Patil et al. (2023) reported the AUCs obtained by their state-of-the-art CNN model at lead times of 18 and 23 months. These values were 0.75, 0.75 and 0.62 for the El Niño, La Niña and neutral classes respectively at 18 months lead time and 0.69, 0.7 and 0.64 at 23 months lead time. For comparison, the macro-averaged AUCs obtained by eSPA and shown in figure 1 are 0.80 and 0.82 for 18 and 24 months lead time respectively. Prior to macro-averaging (i.e. weighting by the class priors), the AUC for each class is 0.84, 0.76 and 0.76 for the El Niño, La Niña and neutral classes respectively at 18 months lead time and 0.85, 0.85 and 0.75 at 24 months lead time. Therefore, eSPA obtains better performance than the model of Patil et al. (2023) for all three classes individually at both of these lead times.

c. Interpretability

One of the big advantages of eSPA is the ability to rigorously examine and interpret the model structure and results, as will be demonstrated in this section. Figure 4 shows plots of the W vector of probabilities, which may be interpreted as a measure of feature importance, for each lead time. It can easily be observed that there is a strong sparsification of the feature space, with only the leading PCs of SST and $\mathrm{d}T/\mathrm{d}z$ making up most of the contribution to the discretisation error. Furthermore, the number of important features, given by $\tilde{D}$, increases with increasing lead time, a reflection of the fact that there is a greater diversity in the patterns used to predict events as lead time increases. Importantly, the leading $\mathrm{d}T/\mathrm{d}z$ PCs are deemed to be important for prediction at all lead times and are allocated a similar weight to the leading SST PCs.

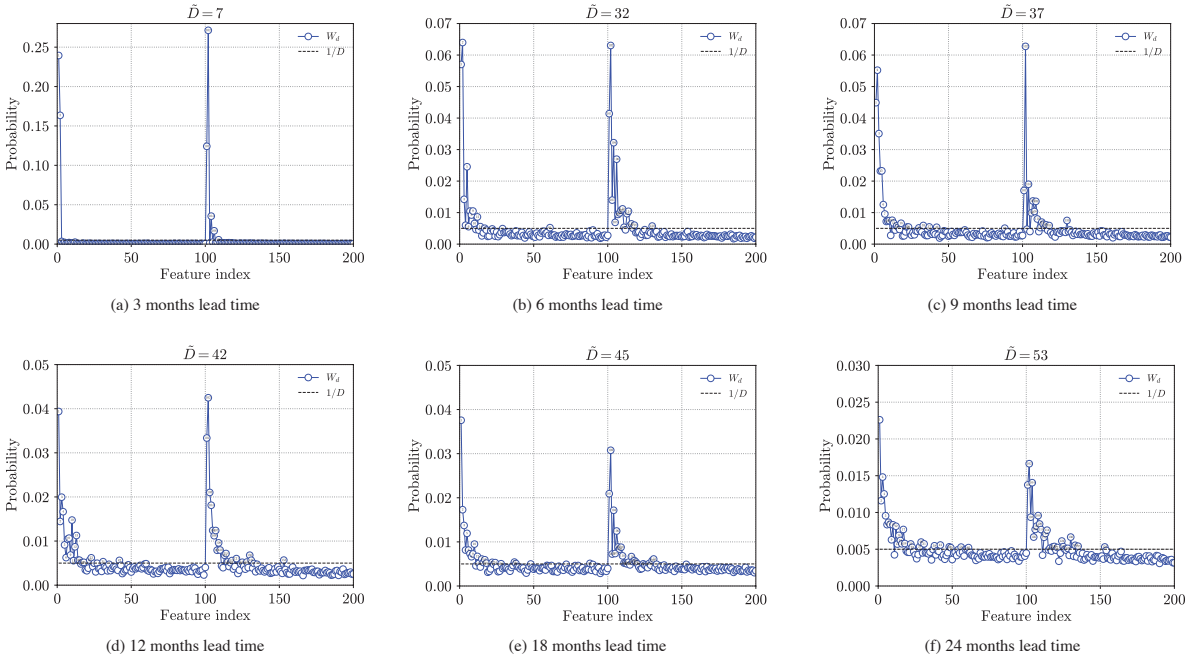
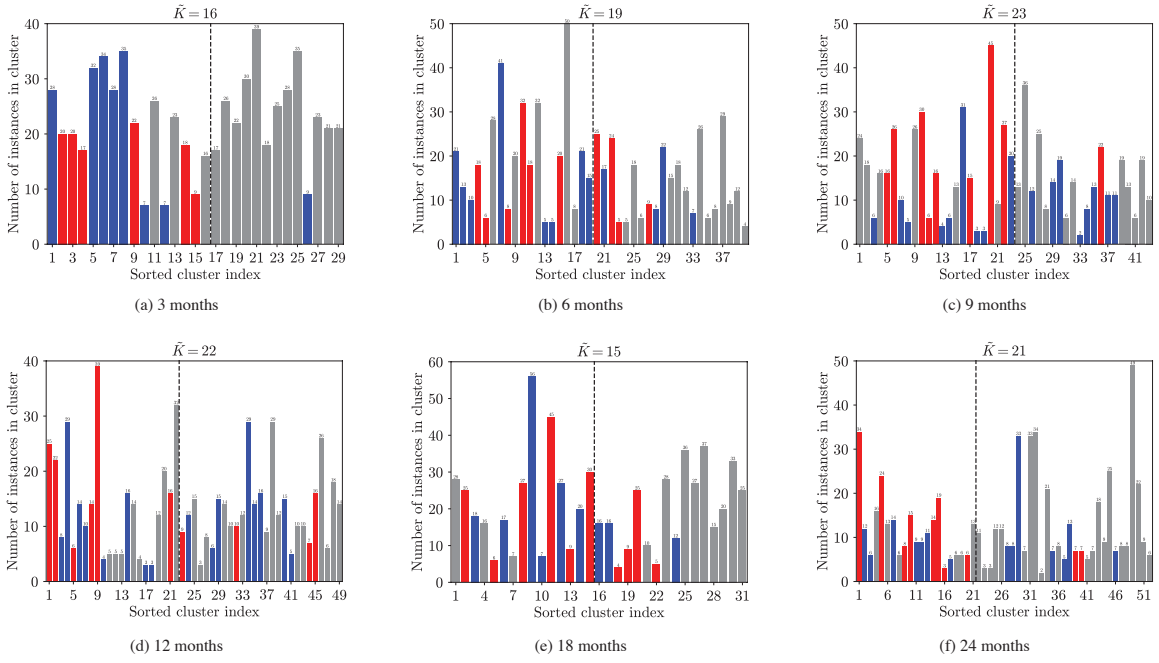


FIG. 4: Feature importance plots for the baseline dataset. The number of features with probability greater than $1/D$ is given in the title of each plot.

In addition to the W vector, the cluster affiliation matrix Γ can be shown to provide useful insight as well. Figure 5 shows occupation plots for the training set, i.e. the number of instances in the training set that are assigned to each cluster. These are sorted by their respective p-value to be in order of most to least significant. Furthermore, by evaluating the highest conditional probability (given by $\text{argmax}(\Lambda_{:,k})$) for each cluster k, the clusters predicting El Niño, La Niña and neutral classes may be distinguished. Of crucial importance is the fact that, for each lead time, the majority of least significant clusters are those associated with the neutral class. This demonstrates that during the grid search for optimal hyperparameters K, ε_E and ε_C the higher than expected value for K that is chosen is done so as not to pollute the most predictive El Niño and La Niña clusters with neutral instances. It can also be observed that the number of significant clusters $\tilde{K}$ is approximately constant with lead time. The number of El Niño clusters is 6, 10, 9, 10, 10 and 10 for each lead time respectively, while the number of La Niña clusters is 8, 12, 17, 16, 9 and 14.

Another important insight gained by inspecting the Γ matrix is to plot the cluster affiliation sequence in time, as shown in figure 6 for the training set and figure 7 for the test set at each lead time (the remaining lead times for the training set can be viewed in the supplementary material).

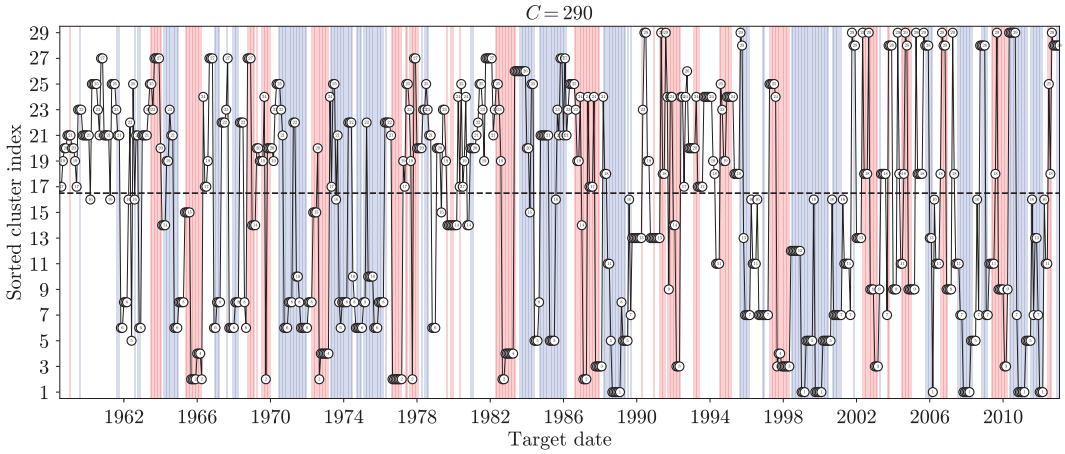


(a) 3 months (b) 6 months (c) 9 months

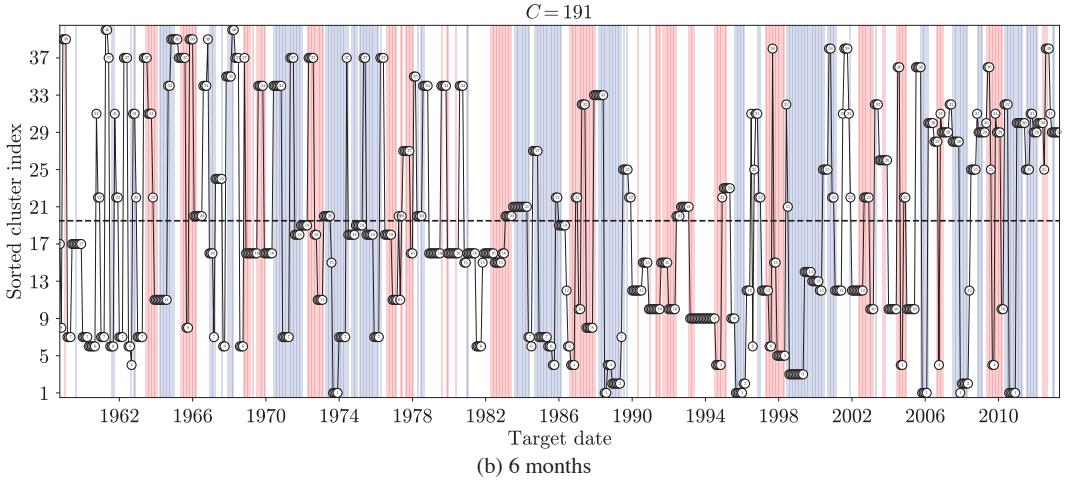
(d) 12 months (e) 18 months (f) 24 months

FIG. 5: Training set cluster occupation plots for the baseline dataset. El Niño clusters are coloured red, La Niña clusters are coloured blue and neutral clusters are coloured grey, while the dashed black line demarcates clusters whose p-value is less than 0.05.

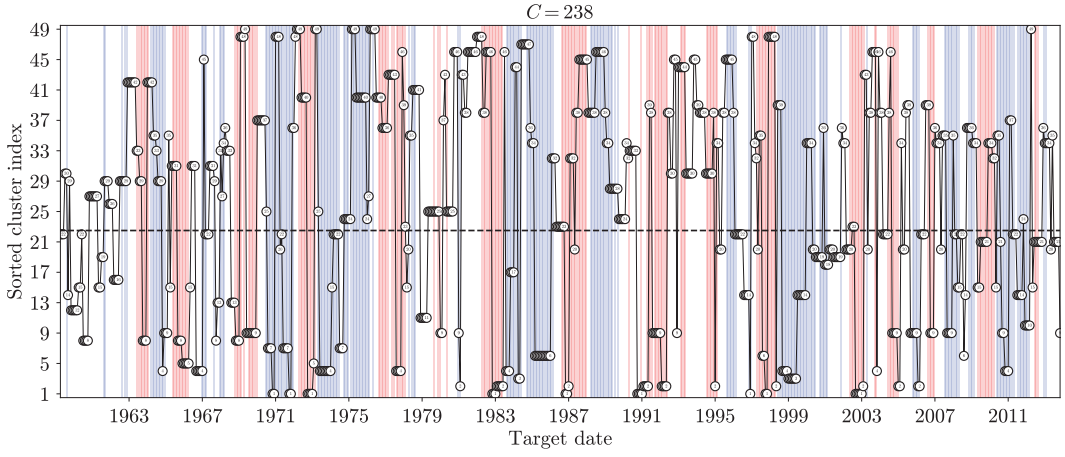
As for the occupation plots, the clusters are sorted from most to least significant p-value. From these plots, the sequence of cluster affiliations occurring n-months prior to an El Niño or La Niña event is able to be determined and can elucidate important insights into how well the model is capturing known ENSO dynamics. Such sequences are also helpful for determining whether a given cluster typically appears earlier or later in the historical period (due to non-stationarity in the feature space) as well as earlier or later in the prediction of a particular event in n months time.



(a) 3 months



(b) 6 months



(c) 12 months

FIG. 6: Training set cluster affiliation plots for the baseline dataset. The red, blue and white background shading indicates an El Niño, La Niña or neutral event occurring on that target date, while the number of switches between clusters is given in the title of each plot.

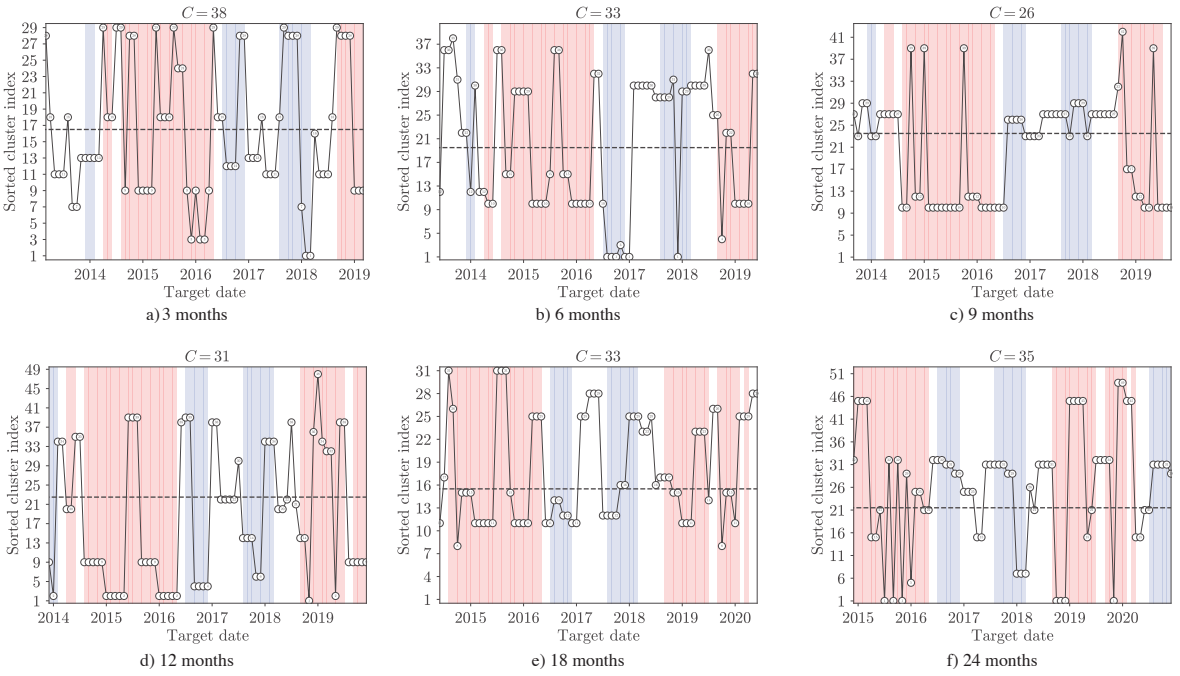


FIG. 7: Test set cluster affiliation plots for the baseline dataset. The red, blue and white background shading indicates an El Niño, La Niña or neutral event occurring on that target date, while the number of switches between clusters is given in the title of each plot.

On the test set the affiliation sequences are useful for determining whether a particular prediction is being made from a highly significant cluster or one that occurred less frequently, or was assigned to a mix of events, over the training set. In general, predictions on the test set that are incorrect are being made from less significant clusters. This provides another way of assessing confidence in the predictions aside from just looking at the conditional probabilities for that cluster contained in the Λ matrix.

To further help in understanding a given affiliation sequence, composite images of SST and $\mathrm{d}T/\mathrm{d}z$ may be generated from the centroids of each cluster, given by the matrix S, as these correspond to a particular value for each PC (after rescaling) that can then be combined with its corresponding EOF to generate a spatial plot of SST or $\mathrm{d}T/\mathrm{d}z$, similar to the methodology used in Wang and Huang (2024). For the sake of brevity, only a few key sequences and their corresponding composite images will be illustrated in detail here; a full set of composites for each lead time is made available in the supplementary material. At shorter lead times the composite images for the most significant clusters correspond to patterns representing mature or emerging/declining El Niño or La Niña states

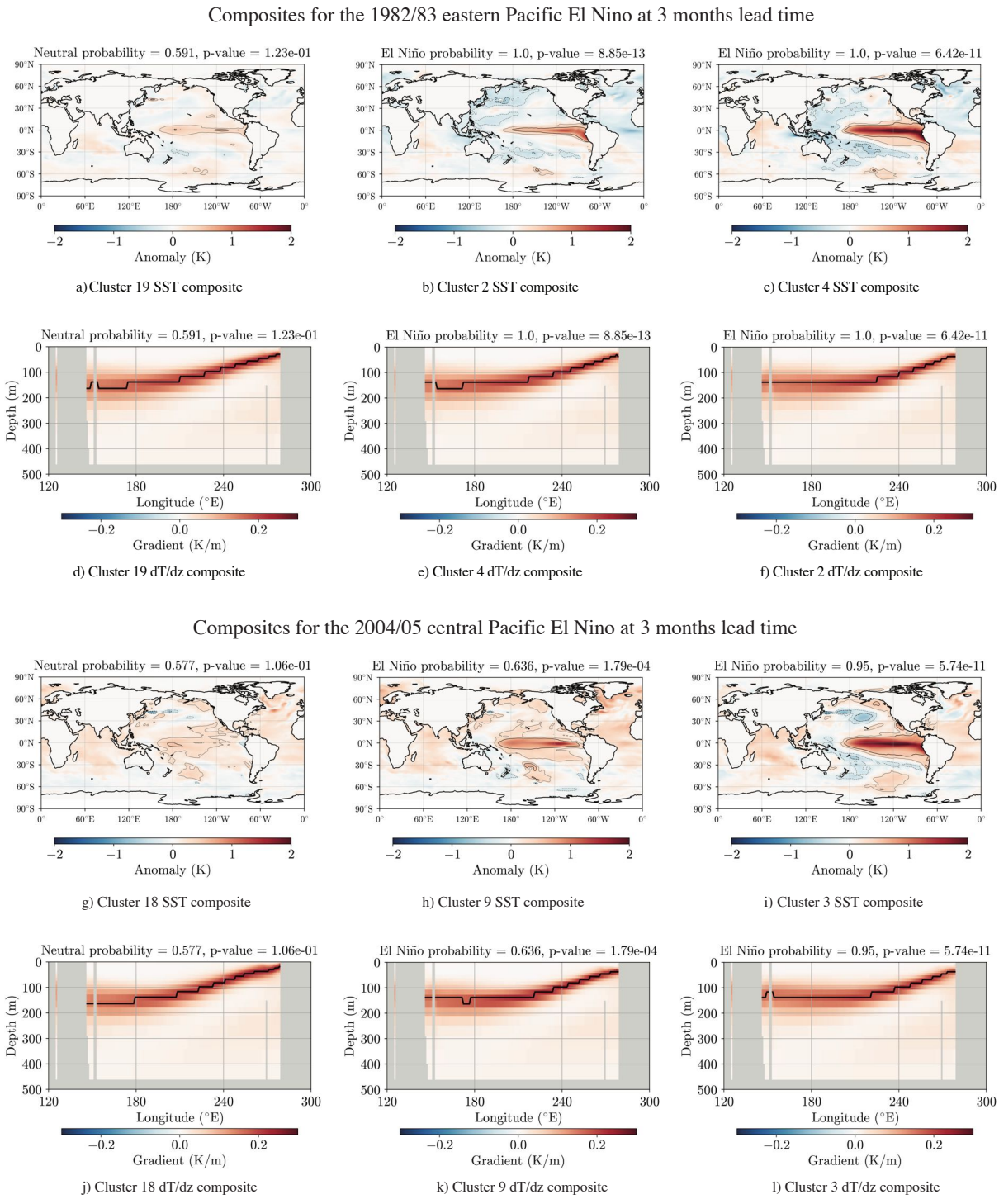


FIG. 8: SST and $\mathrm{d}T/\mathrm{d}z$ composites at 3 months lead time.

(as well as intermediate neutral states), while at longer lead times they correspond to precursor states consisting of extra-tropical anomalies.

Figure 8a-8f shows the SST and dT/dz composites for clusters corresponding to an eastern Pacific El Niño at 3 months lead time, in this case the 1982/83 El Niño event. Prior to the event, cluster 19 is active which exhibits a slightly warm SST anomaly in the central and eastern Pacific and a neutral thermocline position. As the event begins to develop cluster 2 becomes active and the characteristic 'tongue' of warm SST anomaly is present in the eastern Pacific and to a lesser extent the central Pacific. Relative to cluster 19 the thermocline has also begun to shoal. Finally, once the event is well underway cluster 4 becomes active which is representative of a fully developed (eastern Pacific) El Niño with a large warm SST anomaly in the eastern and central Pacific and a fully shoaled thermocline. Given the short lead time of 3 months, this is effectively a persistence forecast; if the Niño3.4 index is well above the threshold of 0.4°C then it is very likely to still be above that threshold in 3 months time.

Figure 8g-8l shows a similar set of composites to figure 8a-8f but for a central Pacific El Niño at 3 months lead time, in this case the 2004/05 El Niño event. Similar to above, the system starts in cluster 18 which is a neutral cluster with a slightly warm SST anomaly in the central-western Pacific and a relatively steep thermocline profile. Towards the onset of the event cluster 9 becomes active, which features a warm SST anomaly in the central Pacific and a shoaled thermocline. After onset cluster 3 becomes active. This cluster is similar to cluster 4 shown in figure 8a-8f but with a weaker SST anomaly off the coast of South America. Relative to cluster 4, the peak SST anomaly is also located further westward along the equator.

Figure 9a-9f shows the SST and dT/dz composites for clusters corresponding to a La Niña at 6 months lead time, in this case the 1988/89 La Niña event. Initially, cluster 33 is active. This cluster has a large warm SST anomaly in the central and eastern Pacific and a shoaled thermocline, since a weak El Niño preceded the 88/89 La Niña in the previous year. More generally, at 6 months lead time there are numerous clusters representing a mature/decaying ENSO phase that predict the opposite phase to occur in 6 months time. This is representative of the observed cases where an El Niño can immediately follow a La Niña and vice versa. Following cluster 33, cluster 1 becomes active which represents a neutral ENSO phase. There are still patches of anomalously warm SST however they do not cross the equator. Furthermore, the thermocline has begun to steepen and a cold SST anomaly is beginning to develop off the coast of South America indicating increased

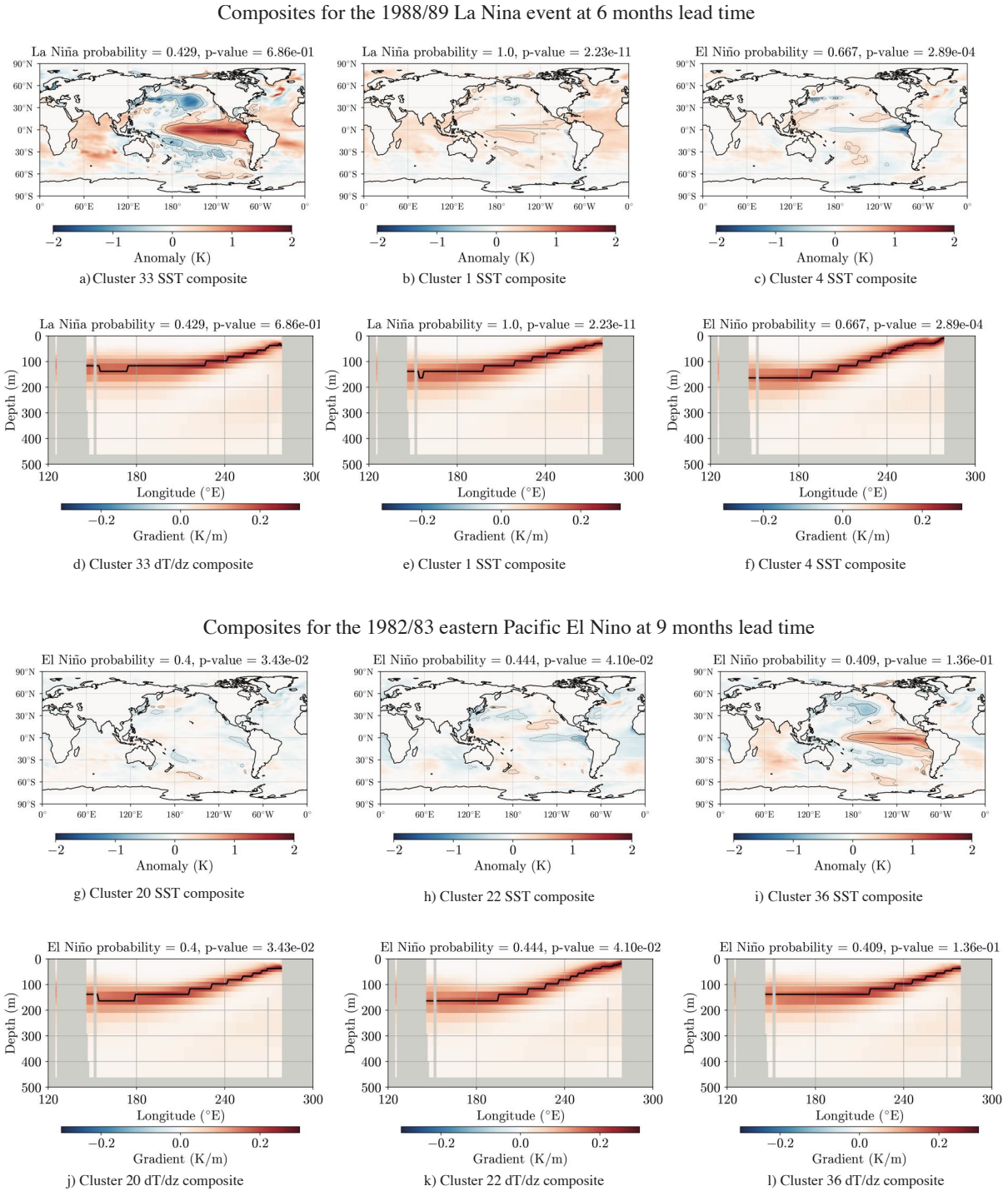


FIG. 9: SST and dT/dz composites at 6 (top) and 9 (bottom) months lead time.

upwelling. Finally, the system moves to cluster 4 which shows the development of a cold SST anomaly along the equator and increased steepening of the thermocline.

Moving to longer lead times, 9g-9l and 10 show composites for the same events shown in figures 8a-8f and 8g-8l but for lead times of 9 months and 12 months respectively. At 9 months lead time, the first cluster indicating the onset of the 1982/83 El Niño, cluster 20, has both SST and $\mathrm{d}T/\mathrm{d}z$ composites that are representative of a neutral phase of ENSO. The system then moves to cluster 22, which features a weak cold SST anomaly in the eastern Pacific along the equator, a weak warm SST anomaly in the central-eastern Pacific north of the equator and a steeper thermocline than cluster 20. Note that the system switches between these two clusters multiple times before moving to cluster 36, which shows the development of a warm SST anomaly along the equator and a shoaling thermocline.

For 12 months lead time prior to the 2004/05 El Niño the system is initially in cluster 23, which features a neutral thermocline, a patch of anomalously cold SST off the coast of South America at the equator, a smaller patch of anomalously warm SST further westward along the equator as well as multiple patches of anomalously warm and cold SSTs at the mid-latitudes. The next cluster is cluster 1, which has a slightly flatter thermocline than cluster 23 as well as a similar patch of anomalously cold SST in the eastern Pacific along the equator, but which is surrounded by equivalently warm SST anomalies in the mid-latitudes and western Pacific. From cluster 1 the system then moves to cluster 2, which has a steeper thermocline profile but also shows the development of anomalously warm SSTs in the equatorial Pacific at the dateline and which extend northeastward towards the coast of North America. Warm SST anomalies are also present in the mid-latitudes east of the dateline. Interestingly, the spatial pattern for the cluster 2 SST composite closely resembles the pattern obtained from the information flow-based causality analysis performed for El Niño Modoki (i.e. a central Pacific El Niño) by Liang et al. (2021).

Note that for many of the improvements to the baseline case described in section 4 below, the clustering is considerably more parsimonious and with improvements in the out-of-sample predictions, i.e. clusters are fewer and more persistent without sacrificing predictive ability. The interested reader is therefore encouraged to view the supplementary material, which contains a complete set of cluster composites for all of the cases presented in section 4 along with the baseline case.

To give an idea of the similarity between clusters a pattern correlation is computed between the SST composites of each cluster, restricted to the Pacific Ocean between ±60° latitude and

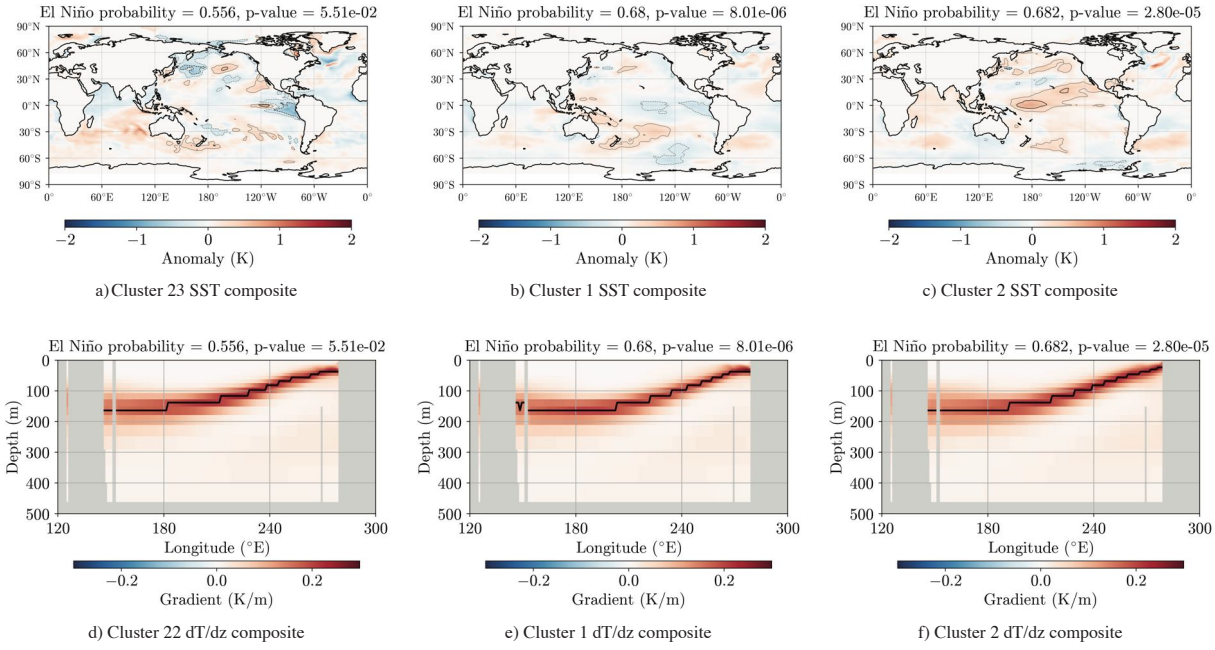


FIG. 10: Composites for the 2004/05 central Pacific El Niño at 12 months lead time.

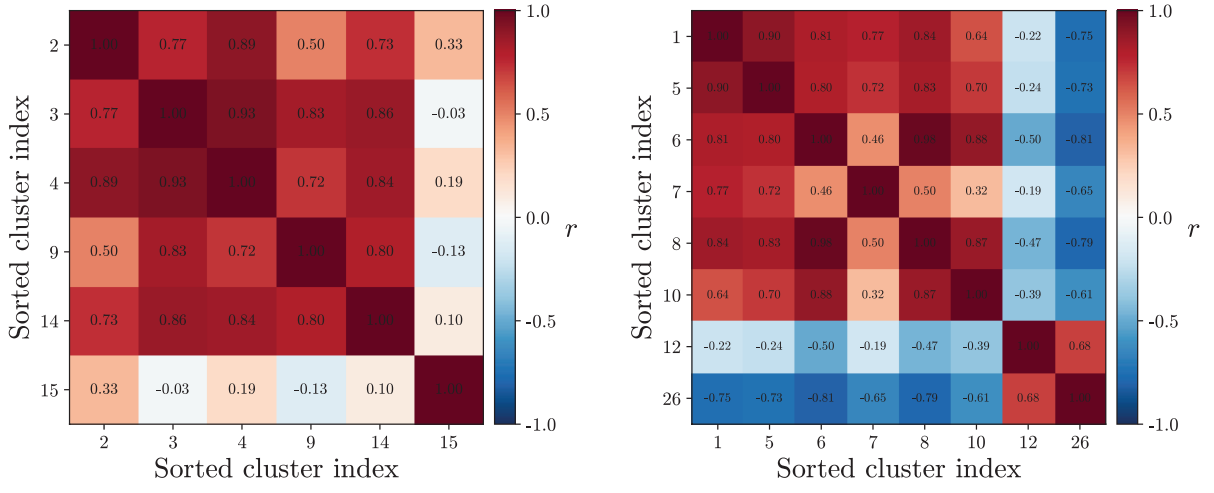
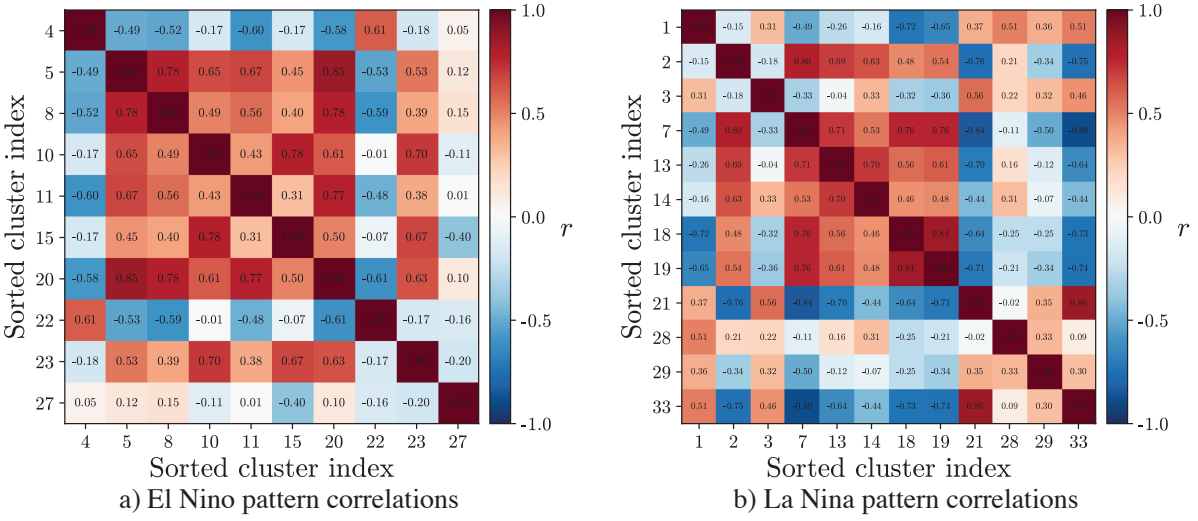


FIG. 11: Pattern correlations for 3 months lead time.

120°-300° longitude. The resulting correlation coefficients are shown in figures 11 and 12 for 3 and 6 months lead time respectively. Similar plots for the remaining lead times are given in the supplementary material. Comparing between lead times, it is clear that there is an increased diversity in the precursor patterns with the highest predictive skill for El Niño and La Niña events



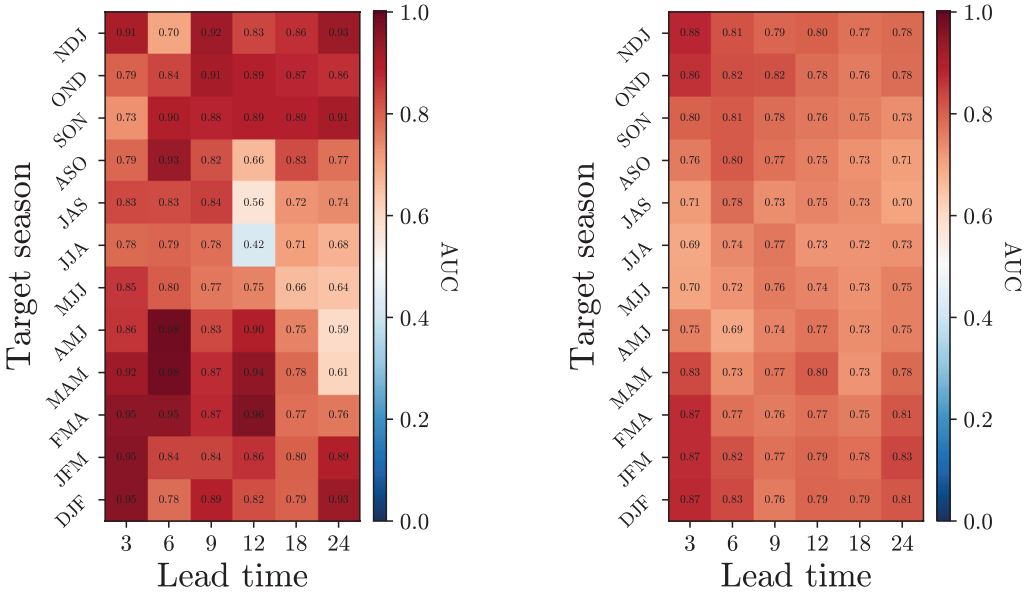
a) El Nino pattern correlations b) La Nina pattern correlations

FIG. 12: Pattern correlations for 6 months lead time.

as lead time increases. This can be summarised by calculating an average pattern correlation coefficient for each lead time of 3, 6, 9, 12, 18 & 24 months, which for the El Niño clusters is 0.554, 0.163, -0.061, -0.020, 0.240 and 0.184 respectively and for the La Niña clusters is 0.191, 0.029, 0.038, 0.082, -0.071 and 0.044. From the high pattern correlations depicted in figures 11 and 12, it is readily apparent that the number of eSPA cluster affiliations at short lead times might be further reduced, however, as lead times increase lower pattern correlations are indicative of the increased diversity of precursors across individual ENSO events.

d. Seasonality of predictions

It is well established that both dynamical and statistical ENSO forecasting systems exhibit a boreal spring predictability barrier arising due to the climatological auto-correlation in tropical Pacific SSTs, in which short lead time forecast skill is at a minimum during the boreal summer, extending to later seasons for longer lead times (Barnston et al. 2012). A similar prediction barrier is also observed in many modern deep learning prediction systems such as the CNN model of Ham et al. (2019). It is therefore interesting to examine the variability in skill of eSPA with each target season to see if both the model and/or problem formulation as a classification task help to reduce this barrier.



a) Seasonality of AUC on test dataset

b) Seasonality of AUC on entire dataset

FIG. 13: Seasonality of eSPA predictions on the baseline dataset.

Figure 13 shows the AUC of the predictions for each target season and lead time, both on the test set as well as the overall dataset. Due to the limited size of the test set, there are only 18 or 19 instances for each 3-month season and for some seasons there is only one or even no instances of a particular class. For this reason, it is difficult to draw any strong conclusions as the small sample sizes introduce a lot of additional variability. To circumvent this issue and increase the sample sizes the same plot is generated for the entire dataset. Despite the model seeing 90% of this data during training, there is sufficient regularisation in eSPA to avoid overfitting and therefore the AUC values on the entire dataset are quite similar to those on the test set; they are 0.82, 0.79, 0.79, 0.79, 0.75 and 0.78 for each lead time respectively. Examining Figure 13b, it can be seen that there is a decrease in skill as measured by AUC for the boreal summer at 3 and 6 months lead time. At 9 months lead time there is a smaller decrease in skill, but the minimum in skill still occurs later in the year (JAS). At 12 months lead time there is a greater decrease in skill again, with the minimum occurring during JJA, while at 18 months lead time the trend is similar to that at 9 months with less variation in skill throughout the year (although lower skill overall). Finally, at 24 months lead time there is a similar variation in skill to that at 12 months, with the minimum in skill occurring during JAS and ASO. From these results, it can be concluded that for all lead times considered there is a

reduction in skill for target seasons falling within the boreal summer, although this reduction does not appear to be as severe as that observed for forecast systems, both dynamical and statistical, that attempt to predict the Niño3.4 index directly.

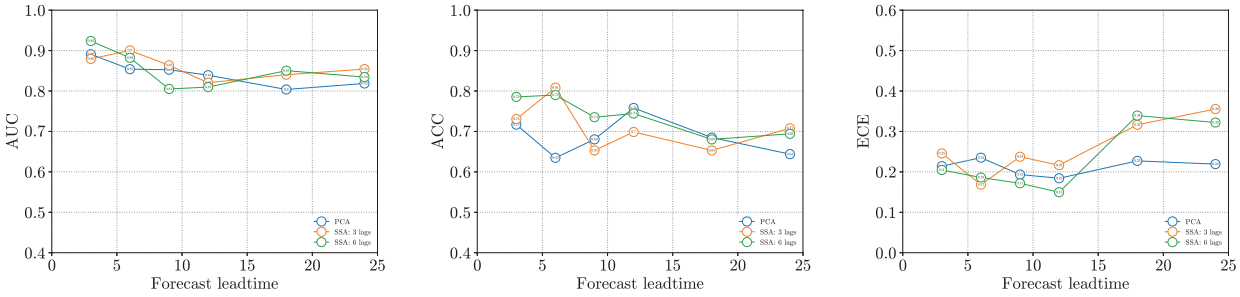
4. Dataset sensitivity

In this section, various modifications to the baseline dataset are made in order to assess how they affect both the predictive power of the obtained eSPA model as well as its interpretability and parsimony.

a. PCA vs. SSA

Rather than just use PCA to decompose the SST and $\mathrm{d}T/\mathrm{d}z$ data into principal components and their corresponding EOFs, a form of multivariate Singular Spectrum Analysis (SSA) can be applied to embed m lags of the data at each time into a lagged covariance matrix, resulting in $m+1$ EOFs for each PC in the decomposition. This can be thought of as an application of Takens' delay embedding theorem (Takens 1981), which states that the dynamics of a system can be reconstructed from a series of observations of a single variable over time (i.e. each PC) provided the embedding dimension is sufficiently large to ensure that the topology of the reconstructed attractor is equivalent to the topology of the original system's attractor. Embedding dimensions of 3 months and 6 months are considered here (referred to as 3 lags and 6 lags hereafter). From an interpretability perspective, embedding lagged instances of the data means that each cluster now contains a sequence of composites that capture the evolution of the system through time within that cluster. Additionally, as will be shown below, it typically results in more parsimonious eSPA models than those obtained from PCA-derived features without sacrificing predictability.

Figure 14 compares the performance metrics on the test set for eSPA models trained using features derived from PCA as well as SSA with a 3-month embedding and SSA with a 6-month embedding. In general, the out-of-sample performance is comparable between PCA and SSA, with SSA having a slight advantage in terms of AUC at longer lead times (although it is also less calibrated at these lead times). Figure 15 compares the various measures of parsimony between the models obtained using PCA and SSA. Across the different lead times, it is clear that using SSA results in the model selecting fewer features as being important (with the exception of SSA with 3



a) AUC b) ACC c) ECE

FIG. 14: Comparison of performance metrics between features generated using PCA vs. SSA.

lags at a lead time of 3 months), as well as a fewer number of total clusters. The ratio of significant clusters to the total number of clusters is generally larger for SSA and at all lead times using SSA results in fewer switches between clusters (with the exception of SSA with 6 lags at a lead time of 3 months). From these results, it is clear that using SSA leads to a more parsimonious eSPA model, even if out-of-sample performance is not improved.

Another advantage of SSA over PCA is the additional information that is conveyed in the composites generated from the cluster centroids due to having the extra lagged data. Figure 16 illustrates this for two clusters taken from the SSA with 6 lags at 3 months lead time by showing the sequence of SST composite images obtained from clusters 2 and 3. Note that these are the clusters that are active for the same 1982/83 eastern Pacific El Niño that is shown in figure 8a-8f. For brevity, the SST composite at 6 months lag has been omitted, along with all of the $\mathrm{d}T/\mathrm{d}z$ composites. The sequence shows the initial onset of the event while the system is in cluster 2 and then the maturation and peak of the event while the system is in cluster 3. Furthermore, the entire progression of the composite images across both clusters is very smooth and continuous and allows for additional interpretability over just using composites obtained from PCA.

b. Sensitivity to labels

The generation of the labels from HadISST observations involves a subjective choice for the threshold to use on the Niño3.4 index above/below which an instance is labelled as El Niño or La Niña. The sensitivity of the out-of-sample performance to this threshold is examined in figure 17a-17b, which compares the test set AUC for both the standard threshold of $\pm 0.4^\circ$ as well as a higher threshold of $\pm 0.8^\circ$. This has the effect of filtering out low amplitude events but also reduces

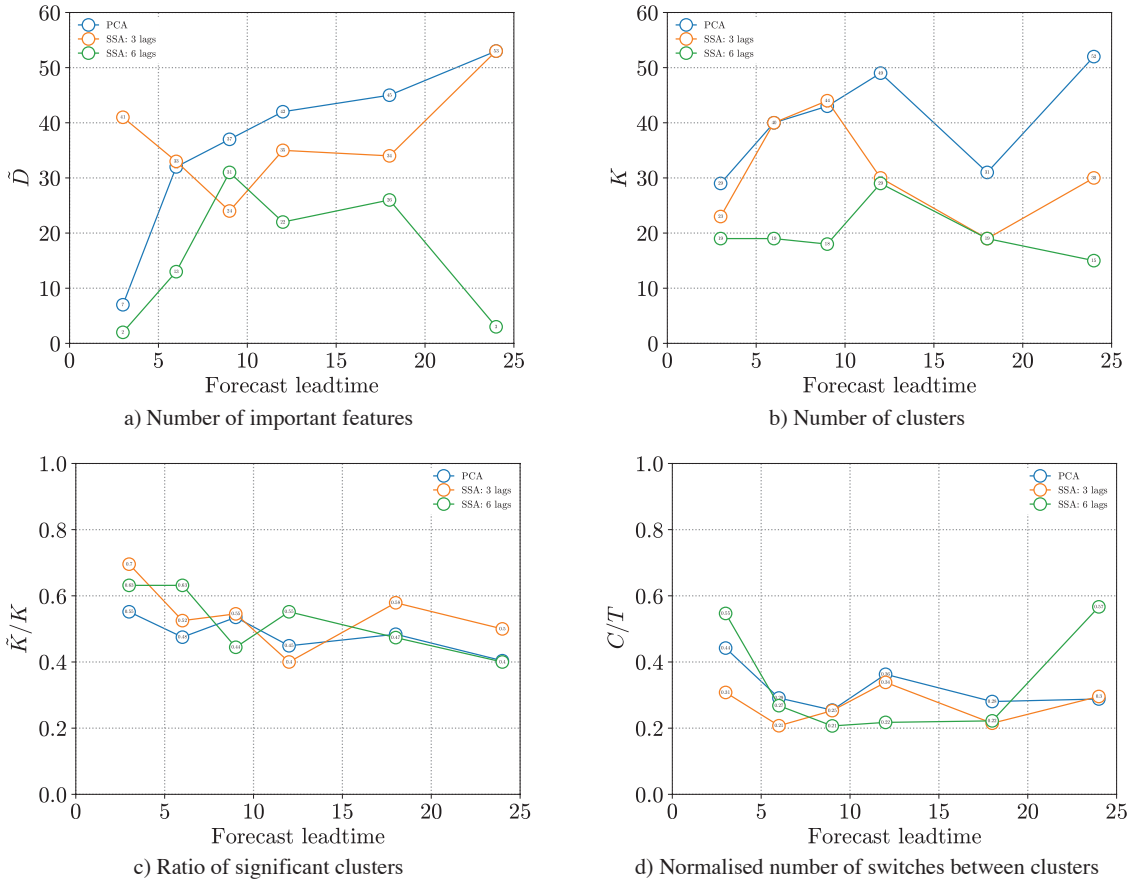


FIG. 15: Measures of parsimony for eSPA models based on PCA- and SSA-derived features.

the class priors for the El Niño and La Niña classes. As a result, the out-of-sample performance is reduced for both PCA- and SSA-derived features due to there being fewer examples to learn from in training.

Another modification that can be made to the labels is to filter out short-lived events. In figure 17c-17d results are shown for both PCA and SSA with 3 lags where the labels have been modified so that any El Niño or La Niña events that do not persist for at least 5 months are labelled as neutral instead (which is the window that NOAA uses when classifying ENSO events). In terms of out-of-sample performance, as measured by AUC on the test set, this has little effect for PCA and a slight decrease in performance for SSA. The effects on model parsimony are given in figure S10 of the supplementary material and are summarised here. Both PCA- and SSA-based models place importance on a larger number of features, while both PCA and SSA use a similar number of total

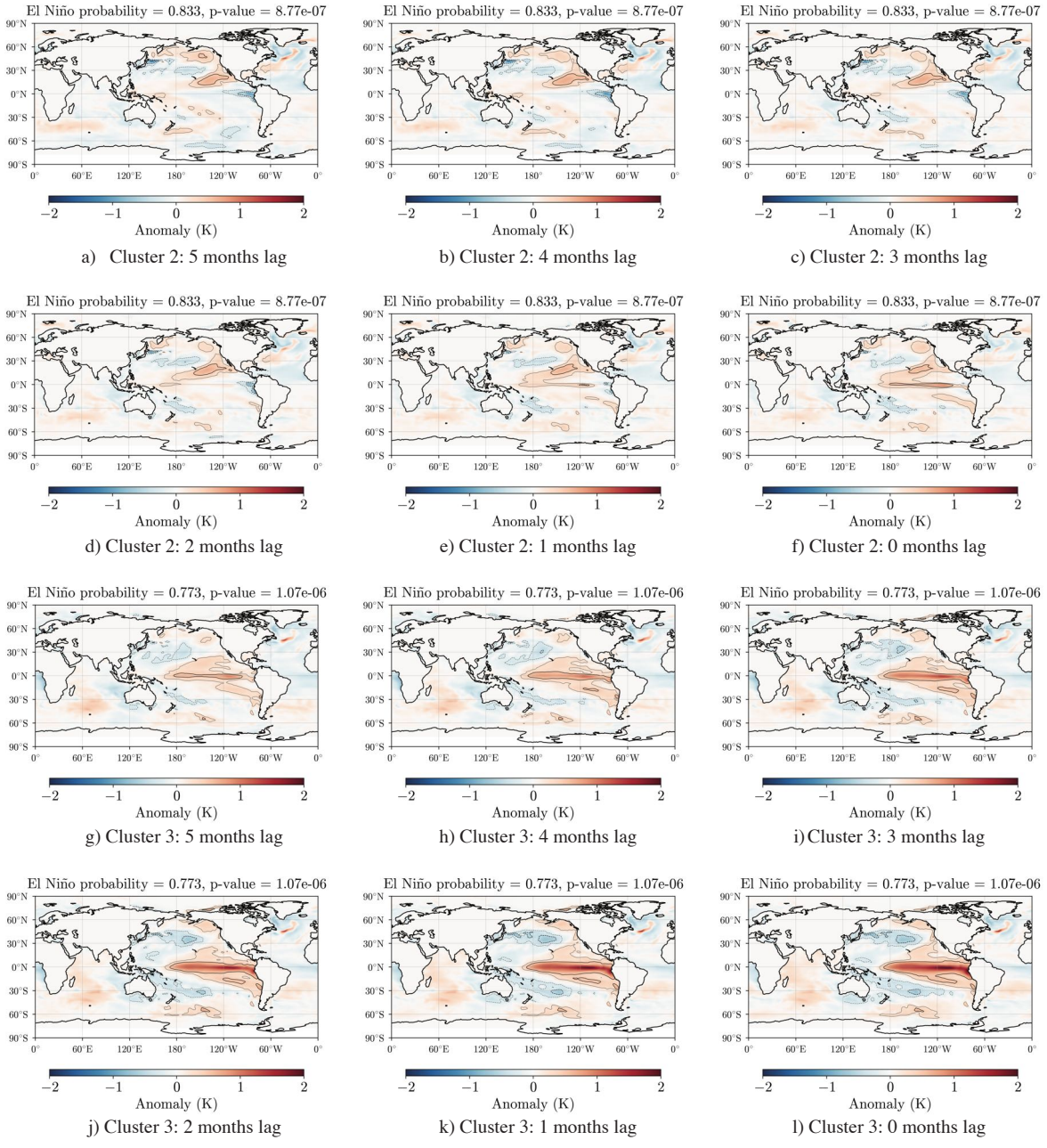


FIG. 16: SST composite images for clusters 2 and 3 for SSA (6-month embedding) at 3 months lead time. Note: the composite at 6 months lag has been omitted for both clusters for brevity.

clusters irrespective of whether the labels are filtered or unfiltered (with the exception of PCA at 18 months lead time and SSA at 24 months lead time). The ratio of significant clusters is slightly reduced for both PCA and SSA and the number of switches between clusters is slightly increased.

Based on these findings it appears unnecessary to filter out short-lived events, however it may still be useful for certain applications that are targeting a particular timescale.

c. Sensitivity to amount of training data

Finally, in this section the sensitivity to the amount of training data used is explored for both PCA and SSA with 3 lags. This is of interest for multiple reasons. Firstly, if it can be shown that eSPA does not overfit when there is even less training data than is provided by the full ACCESS-OM2 dataset then this allows for the possibility of instead using reanalysis products that only cover the more recent historical period such as the NCEP Global Ocean Data Assimilation System (Behringer et al. 1998). Furthermore, it is not uncommon for the first few years of an ocean model dataset to be unrealistic due to the spin-up used to initialise the ocean state. For example, the ACCESS-OM2 model uses atmospheric forcing taken from the JRA-55 reanalysis over the period of 1958 to 2018. This forcing is also used during the spin-up of the model, meaning that the system undergoes a large shock every 60 years when the forcing jumps from 2018 back to 1958, the result of which is that the ocean heat content can take up to 20 years to stabilise to the shock each time.

Secondly, ENSO can be thought of as being superimposed upon a background regime given by the phase of the Interdecadal Pacific Oscillation (IPO). Positive phases of the IPO are characterised by a warmer-than-average tropical Pacific and cooler-than-average northern Pacific, while negative phases are characterised by cooler tropics and warmer northern regions. The implications of this are that ENSO variability, as well as predictability, changes with the phase of the IPO. For example, following the phase change from negative to positive in the mid-1970s, El Niño events increased in frequency but also became easier to predict (O'Kane et al. 2014). Thirdly, the climate system as a whole is highly non-stationary (at least over the observed period), even in the absence of anthropogenic forcing, and as such recent data is more relevant to predicting future conditions than data occurring earlier in time. In data science and machine learning this is known as concept drift, where the statistical properties of the target variable(s), as well as the causal relations between predictors and targets, change over time in unforeseen ways.

The combination of all three of these phenomena can also be observed in the cluster affiliation plots for the training set given in figure 6, where many of the clusters that appear prior to ~ 1980 do not appear afterwards and vice versa. This indicates that reducing the size of the training set by

removing data from the start of the dataset may result in improved, or at least comparable, out-of-sample performance on the final 10% of the dataset. Figure 17e-17f shows the out-of-sample AUC on the final 10% of the dataset when the first 10%, 20%, ..., 50% of the dataset has been removed. Note that for these plots the test set size is kept proportional to the original dataset (i.e. it is the final 10% of the truncated dataset), while in figure 17g-17h the test set is kept fixed at 10% of the original dataset size. For the proportional test set size, there is an improvement in out-of-sample performance relative to the baseline dataset for nearly all lead times and reductions in training data for both PCA and SSA. For PCA the best overall performance is obtained when 50% of the dataset is retained, particularly at longer lead times, while for SSA the best overall performance is obtained for 60% and 70% of the original dataset. In fact, the performance is only worse than the baseline dataset for a lead time of 18 months in a handful of cases.

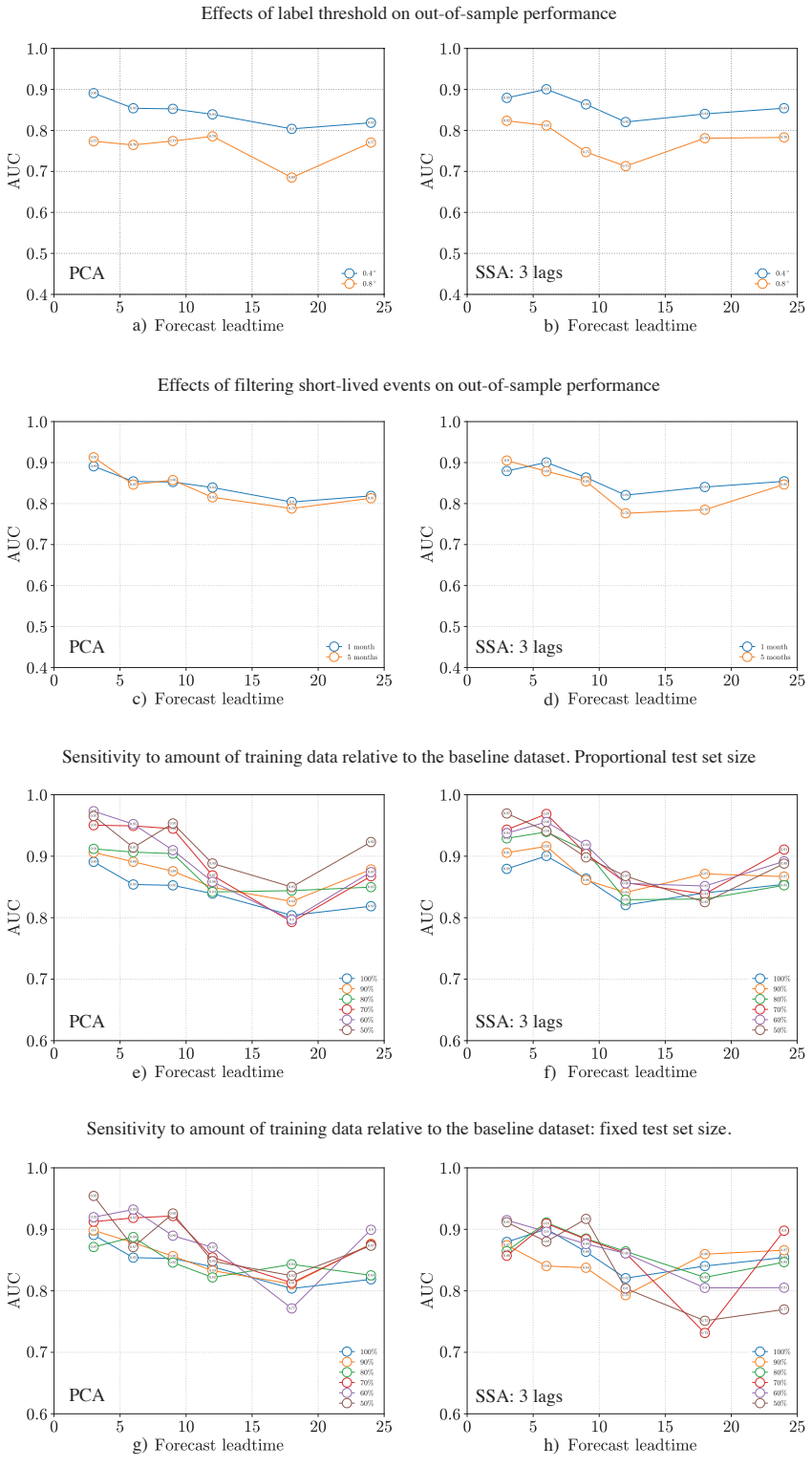


FIG. 17: AUC vs. lead time for various modifications to the baseline dataset.

It is important to note however that the reduction in test set size means that the test set does not contain the same instances across different overall dataset sizes. To explore how this affects the predictions, figure 17g-17h shows the out-of-sample AUC when the test set is kept fixed at 10% of the original dataset size. Compared to the baseline dataset there are still improvements observed, although not quite as substantial as before. Nevertheless, for PCA retaining 50% or 60% of the original dataset still results in a good overall improvement in performance, while for SSA retaining 60% or 70% of the original dataset gives good improvement but less so at the longer lead times.

Figures S11 and S12 in the supplementary material show the same measures of parsimony presented in figure 15 but for the various reductions in dataset size (with the proportional test set). As a general rule for SSA and to a lesser extent PCA, initial reductions in the dataset size result in fewer important features being used in the model, however further reductions past 70% result in models containing a greater number of important features. For PCA nearly all of the models trained on reduced amounts of data used fewer total clusters and, at shorter lead times in particular, contained a larger number of significant clusters and had fewer switches on average between clusters. The same trends are also observed for SSA in general with some exceptions; the baseline dataset has a relatively larger number of significant clusters at longer lead times and differences in the average number of switches between clusters are less clear-cut.

Overall the results in this section suggest that good performance can be obtained for eSPA models using substantially less data than is available in the full ACCESS-OM2 dataset. Note that further improvements would likely be expected if the number of features was also reduced, for example features 150-200 (i.e. the final 50 $\mathrm{d}T/\mathrm{d}z$ PCs) were rarely deemed to be important in the W vector which would justify their removal from the final model. The demonstrated improvements with reductions in training data suggest that perhaps multiple eSPA models could be fitted to different periods. Initially, this could be done according to the different phases of the IPO, however, a more satisfactory solution would be to incorporate some form of temporal regularisation for switching between models trained on different metastable states (Horenko 2010; De Wiljes et al. 2014). Differences in the model structure obtained for each metastable state could then be used to probe how the dynamics of ENSO have changed over the observed period. Such investigations will be performed in future work.

5. Conclusions

This paper has applied the entropy-optimal Scalable Probabilistic Approximation algorithm to systematically classify ENSO variability over the period of 1958-2018 in a resimulated ocean model dataset. A Long Short-Term Memory classifier is also included as a benchmark representative of state-of-the-art deep learning methods *for the same dataset*. For lead times ranging from 3 months to 24 months, both classifiers are trained to predict the phase of ENSO as one of three possible classes (El Niño, La Niña or neutral) based on the first 100 principal components taken from an empirical orthogonal function analysis of global sea surface temperature anomalies as well as the vertical derivative of subsurface temperature at the equator, both derived from the resimulated ocean model.

Relative to the LSTM model, eSPA exhibited substantially better out-of-sample performance in terms of area under the ROC curve for all lead times, better mean accuracy at all lead times (substantially better for lead times greater than or equal to 12 months) and better expected calibration error for lead times greater than or equal to 12 months (worse for lead times less than 12 months). In most cases, the misclassifications by eSPA on the test set were between the neutral class and El Niño or La Niña. The predicted conditional probability for each instance in the test set was also shown to be correlated with the amplitude of the underlying Niño3.4 index from which the labels were generated. In contrast, the LSTM model struggled to correctly predict the neutral class at shorter lead times while at longer lead times it struggled to predict classes other than neutral. The LSTM predictions also exhibited little variation in the probability assigned to each class. In terms of computational cost, it was demonstrated that an eSPA model can be completely trained (i.e. converged to a local minimum) for every 30-220 epochs of training the most competitive LSTM model on the same dataset. Note that since the LSTM model is trained using first-order, gradient-based optimisation it cannot be shown to have completely converged to a local minimum of the loss function. However, in practice approximately 10,000-100,000 epochs were required to observe no further improvement in the loss when training the best possible model (which still exhibited substantially worse performance as described above).

Another major advantage of eSPA is the ease at which its predictions can be thoroughly interrogated and interpreted. By examining the feature importance vector W, it was shown that eSPA induces a strong (but approximate) sparsification of the feature space, with only the leading PCs of

SST and $\mathrm{d}T/\mathrm{d}z$ making up most of the contribution to the discretisation error. By looking at the significance of each of the obtained clusters (as determined by Fisher's exact hypothesis test), the majority of least significant clusters were those associated with the neutral class while the number of significant clusters (at the 5% level) was found to be approximately constant with lead time. Furthermore, incorrect predictions on the test set appear to be made from less significant clusters in general. This provides an additional measure of confidence in the predictions being made aside from the conditional probabilities coming from the Λ matrix.

A novel contribution presented in this paper has been the generation of composite images for each of the clusters by combining the PCs for each of the cluster centroids with their corresponding EOFs. From these reconstructions, it was seen that at shorter lead times the composite images for the most significant clusters correspond to patterns representing mature or emerging/declining El Niño or La Niña states (as well as intermediate neutral states), while at longer lead times they correspond to precursor states consisting of extra-tropical anomalies. Computing the pattern correlations between composites, an increased diversity was observed in the precursor patterns with the highest predictive skill for El Niño and La Niña events with increasing lead time. For all lead times considered, decomposing the predictions into their target seasons showed that, although there is a slight reduction in skill for target seasons falling within the boreal summer, this reduction does not appear to be as severe as that observed for forecast systems that attempt to predict the Niño3.4 index directly. Finally, further modifications to the baseline ACCESS-OM2 dataset were explored showing that improvements can be made in the parsimony of the trained eSPA model without sacrificing predictive power. These modifications included using PCs generated from a multivariate Singular Spectrum Analysis rather than PCA, as well as reducing the amount of training data by only including more recent observations as a means of circumventing concept drift, of which the latter also resulted in improved out-of-sample performance.

Given the promising results presented in this paper, there are numerous options for extending the current method and analysis in future work, of which just a few are mentioned here. Firstly, as mentioned already in section 4, the demonstrated improvements with reductions in training data suggest that multiple eSPA models could be fitted to different periods of the historical record, with some form of temporal regularisation governing the switching between models trained on different metastable states. Secondly, there is the potential to develop eSPA, as well as its analogous

regression formulation SPARTAn, into a proper ENSO forecast system. A proof-of-concept for this has already been presented in Horenko et al. (2023). This would then be thoroughly assessed using hindcasts in the same way as conventional dynamical and statistical forecast systems (Barnston et al. 2012). Finally, it would be interesting to apply eSPA to daily atmospheric data to look at classifying climate variability on intraseasonal time scales, for example by predicting the Madden-Julian oscillation.

Acknowledgments. The authors wish to acknowledge the Consortium for Ocean-Sea Ice Modelling in Australia (COSIMA; `http://www.cosima.org.au`) for providing access to the ACCESS-OM2 model outputs. MG would also like to thank Dylan Harries for providing code for calculating the empirical orthogonal function decomposition for gridded climate data.

Data availability statement. The HadISST dataset is available at `https://www.metoffice.gov.uk/hadobs/hadisst/`. The ACCESS-OM2 model configuration used in this paper is available at `https://github.com/hakaseh/1deg_jra55_iaf/tree/omip2`, while further information on ACCESS-OM2 can be found at `https://cosima.org.au/index.php/models/access-om2/`. The eSPA source code can be made available via email request to the authors. The supplementary material is available at `https://zenodo.org/records/10582420`.

References

Barnston, A. G., M. K. Tippett, M. L. L'Heureux, S. Li, and D. G. DeWitt, 2012: Skill of Real-Time Seasonal ENSO Model Predictions During 2002–11: Is Our Capability Increasing? *Bulletin of the American Meteorological Society*, **93 (5)**, ES48–ES50, https://doi.org/10.1175/BAMS-D-11-00111.2, URL https://journals.ametsoc.org/doi/10.1175/BAMS-D-11-00111.2.

Behringer, D. W., M. Ji, and A. Leetmaa, 1998: An improved coupled model for enso prediction and implications for ocean initialization. part i: The ocean data assimilation system. *Monthly Weather Review*, **126 (4)**, 1013 – 1021.

Bjerknes, J., 1969: ATMOSPHERIC TELECONNECTIONS FROM THE EQUATORIAL PACIFIC [1]. *Monthly Weather Review*, **97 (3)**, 163–172, https://doi.org/10.1175/1520-0493(1969)097⟨0163:ATFTEP⟩2.3.CO;2, URL http://journals.ametsoc.org/doi/10.1175/1520-0493(1969)097⟨0163:ATFTEP⟩2.3.CO;2.

Chen, N., F. Gilani, and J. Harlim, 2021: A Bayesian Machine Learning Algorithm for Predicting ENSO Using Short Observational Time Series. *Geophysical Research Letters*, **48 (17)**, e2021GL093 704, https://doi.org/10.1029/2021GL093704, URL https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021GL093704.

Chen, N., C. Su, S. Wu, and Y. Wang, 2023: El Niño Index Prediction Based on Deep Learning with STL Decomposition. *Journal of Marine Science and Engineering*, **11 (8)**, 1529, https://doi.org/10.3390/jmse11081529, URL https://www.mdpi.com/2077-1312/11/8/1529.

De Wiljes, J., L. Putzig, and I. Horenko, 2014: Discrete nonhomogeneous and nonstationary logistic and Markov regression models for spatiotemporal data with unresolved external influences. *Communications in Applied Mathematics and Computational Science*, **9 (1)**, 1–46, https://doi.org/10.2140/camcos.2014.9.1, URL http://msp.org/camcos/2014/9-1/p01.xhtml.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, **9 (5)**, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, URL https://gmd.copernicus.org/articles/9/1937/2016/.

Gao, C., L. Zhou, and R. Zhang, 2023: A Transformer-Based Deep Learning Model for Successful Predictions of the 2021 Second-Year La Niña Condition. *Geophysical Research Letters*, **50 (12)**, e2023GL104 034, https://doi.org/10.1029/2023GL104034, URL https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL104034.

Gerber, S., L. Pospisil, M. Navandar, and I. Horenko, 2020: Low-cost scalable discretization, prediction, and feature selection for complex systems. *SCIENCE ADVANCES*.

Giese, B. S., and S. Ray, 2011: El Niño variability in simple ocean data assimilation (SODA), 1871–2008. *Journal of Geophysical Research*, **116 (C2)**, C02 024, https://doi.org/10.1029/2010JC006695, URL http://doi.wiley.com/10.1029/2010JC006695.

Ham, Y.-G., J.-H. Kim, E.-S. Kim, and K.-W. On, 2021: Unified deep learning model for El Niño/Southern Oscillation forecasts by incorporating seasonality in climate data. *Science Bulletin*, **66 (13)**, 1358–1366, https://doi.org/10.1016/j.scib.2021.03.009, URL https://linkinghub.elsevier.com/retrieve/pii/S2095927321002243.

Ham, Y.-G., J.-H. Kim, and J.-J. Luo, 2019: Deep learning for multi-year ENSO forecasts. *Nature*, **573 (7775)**, 568–572.

Horenko, I., 2010: On the Identification of Nonstationary Factor Models and Their Application to Atmospheric Data Analysis. *Journal of the Atmospheric Sciences*, **67 (5)**, 1559–

1574, https://doi.org/10.1175/2010JAS3271.1, URL https://journals.ametsoc.org/doi/10.1175/2010JAS3271.1.

Horenko, I., 2020: On a Scalable Entropic Breaching of the Overfitting Barrier for Small Data Problems in Machine Learning. *Neural Computation*, **32 (8)**, 1563–1579, https://doi.org/10.1162/neco_a_01296, URL https://doi.org/10.1162/neco_a_01296.

Horenko, I., E. Vecchi, J. Kardoš, A. Wächter, O. Schenk, T. J. O'Kane, P. Gagliardini, and S. Gerber, 2023: On cheap entropy-sparsified regression learning. *Proceedings of the National Academy of Sciences*, **120 (1)**, e2214972 120, https://doi.org/10.1073/pnas.2214972120, URL https://pnas.org/doi/10.1073/pnas.2214972120.

Jajcay, N., 2018: Synchronization and causality across time scales in El Niño Southern Oscillation.

Jaynes, E. T., 1957a: Information theory and statistical mechanics. *Phys. Rev.*, **106**, 620–630, https://doi.org/10.1103/PhysRev.106.620, URL https://link.aps.org/doi/10.1103/PhysRev.106.620.

Jaynes, E. T., 1957b: Information theory and statistical mechanics. ii. *Phys. Rev.*, **108**, 171–190, https://doi.org/10.1103/PhysRev.108.171, URL https://link.aps.org/doi/10.1103/PhysRev.108.171.

Jensen, J. L. W. V., 1906: Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, **30 (none)**, 175 – 193, https://doi.org/10.1007/BF02418571, URL https://doi.org/10.1007/BF02418571.

Jin, E. K., and Coauthors, 2008: Current status of ENSO prediction skill in coupled ocean–atmosphere models. *Climate Dynamics*, **31 (6)**, 647–664, https://doi.org/10.1007/s00382-008-0397-3, URL http://link.springer.com/10.1007/s00382-008-0397-3.

Jin, F.-F., 1997: An Equatorial Ocean Recharge Paradigm for ENSO. Part I: Conceptual Model. *Journal of the Atmospheric Sciences*, **54 (7)**, 811–829, https://doi.org/10.1175/1520-0469(1997)054⟨0811:AEORPF⟩2.0.CO;2, URL http://journals.ametsoc.org/doi/10.1175/1520-0469(1997)054⟨0811:AEORPF⟩2.0.CO;2.

Kim, J., M. Kwon, S.-D. Kim, J.-S. Kug, J.-G. Ryu, and J. Kim, 2022: Spatiotemporal neural network with attention mechanism for El Niño forecasts. *Scientific Reports*,

12 (1), 7204, https://doi.org/10.1038/s41598-022-10839-z, URL https://www.nature.com/articles/s41598-022-10839-z.

Kiss, A. E., and Coauthors, 2020: Access-om2 v1.0: a global ocean–sea ice model at three resolutions. *Geoscientific Model Development*, **13 (2)**, 401–442.

Liang, X. S., F. Xu, Y. Rong, R. Zhang, X. Tang, and F. Zhang, 2021: El Niño Modoki can be mostly predicted more than 10 years ahead of time. *Scientific Reports*, **11 (1)**, 17 860, https://doi.org/10.1038/s41598-021-97111-y, URL https://www.nature.com/articles/s41598-021-97111-y.

Neelin, J. D., D. S. Battisti, A. C. Hirst, F. Jin, Y. Wakata, T. Yamagata, and S. E. Zebiak, 1998: ENSO theory. *Journal of Geophysical Research: Oceans*, **103 (C7)**, 14 261–14 290, https://doi.org/10.1029/97JC03424, URL https://agupubs.onlinelibrary.wiley.com/doi/10.1029/97JC03424.

O'Kane, T. J., R. J. Matear, M. A. Chamberlain, and P. R. Oke, 2014: Enso regimes and the late 1970's climate shift: The role of synoptic weather and south pacific ocean spiciness. *Journal of Computational Physics*, **271**, 19–38.

Patil, K. R., T. Doi, V. R. Jayanthi, and S. Behera, 2023: Deep learning for skillful long-lead ENSO forecasts. *Frontiers in Climate*, **4**, 1058 677, https://doi.org/10.3389/fclim.2022.1058677, URL https://www.frontiersin.org/articles/10.3389/fclim.2022.1058677/full.

Qiao, S., C. Zhang, X. Zhang, K. Zhang, H. Shi, S. Li, and H. Wei, 2023: Tendency-and-attention-informed deep learning for ENSO forecasts. *Climate Dynamics*, **61 (11-12)**, 5271–5286, https://doi.org/10.1007/s00382-023-06854-z, URL https://link.springer.com/10.1007/s00382-023-06854-z.

Rasmusson, E. M., and T. H. Carpenter, 1982: Variations in tropical sea surface temperature and surface wind fields associated with the southern oscillation/el niño. *Monthly Weather Review*, **110 (5)**, 354 – 384, https://doi.org/https://doi.org/10.1175/1520-0493(1982)110⟨0354:VITSST⟩2.0.CO;2, URL https://journals.ametsoc.org/view/journals/mwre/110/5/1520-0493_1982_110_0354_vitsst_2_0_co_2.xml.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice,

and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research: Atmospheres*, **108 (D14)**, https://doi.org/https://doi.org/10.1029/2002JD002670, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002JD002670, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2002JD002670.

Takens, F., 1981: Detecting strange attractors in turbulence. *Dynamical Systems and Turbulence, Warwick 1980*, D. Rand, and L.-S. Young, Eds., Springer Berlin Heidelberg, Berlin, Heidelberg, 366–381.

Taylor, J., and M. Feng, 2022: A deep learning model for forecasting global monthly mean sea surface temperature anomalies. *Frontiers in Climate*, **4**, 932 932, https://doi.org/10.3389/fclim.2022.932932, URL https://www.frontiersin.org/articles/10.3389/fclim.2022.932932/full.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An Overview of CMIP5 and the Experiment Design. *Bulletin of the American Meteorological Society*, **93 (4)**, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1, URL https://journals.ametsoc.org/doi/10.1175/BAMS-D-11-00094.1.

Vecchi, E., L. Pospíšil, S. Albrecht, T. J. O'Kane, and I. Horenko, 2022: eSPA+: Scalable Entropy-Optimal Machine Learning Classification for Small Data Problems. *Neural Computation*, **34 (5)**, 1220–1255, https://doi.org/10.1162/neco_a_01490, URL https://doi.org/10.1162/neco_a_01490.

Wang, T., and P. Huang, 2024: Superiority of a Convolutional Neural Network Model over Dynamical Models in Predicting Central Pacific ENSO. *Advances in Atmospheric Sciences*, **41 (1)**, 141–154, https://doi.org/10.1007/s00376-023-3001-1, URL https://link.springer.com/10.1007/s00376-023-3001-1.

Wang, Y., Y. Zhang, and G.-G. Wang, 2023: Forecasting ENSO using convolutional LSTM network with improved attention mechanism and models recombined by genetic algorithm in CMIP5/6. *Information Sciences*, **642**, 119 106, https://doi.org/10.1016/j.ins.2023.119106, URL https://linkinghub.elsevier.com/retrieve/pii/S0020025523006916.

Zhou, L., and R.-H. Zhang, 2023: A self-attention–based neural network for three-dimensional multivariate modeling and its skillful ENSO predictions. *Science Advances*, **9 (10)**, eadf2827, https://doi.org/10.1126/sciadv.adf2827, URL https://www.science.org/doi/10.1126/sciadv.adf2827.