Multidecadal Sea Level Prediction Using Neural Networks and Spectral Clustering on Climate Model Large Ensembles and Satellite Altimeter Data

SAUMYA SINHA,[a] JOHN FASULLO,[b] R. STEVEN NEREM,[a] AND CLAIRE MONTELEONI[a,c]

[a] *University of Colorado, Boulder, Boulder, Colorado*
[b] *National Center for Atmospheric Research, Boulder, Colorado*
[c] *INRIA, Paris, France*

(Manuscript received 3 October 2023, in final form 13 July 2024, accepted 1 October 2024)

ABSTRACT: Sea surface height observations provided by satellite altimetry since 1993 show a rising rate (3.4 mm yr^{-1}) for global mean sea level. While on average, sea level has risen 10 cm over the last 30 years, there is considerable regional variation in the sea level change. Through this work, we predict sea level trends 30 years into the future at a 2° spatial resolution and investigate the future patterns of the sea level change. We show the potential of machine learning (ML) in this challenging application of long-term sea level forecasting over the global ocean. Our approach incorporates sea level data from both altimeter observations and climate model simulations. We develop a supervised learning framework using fully connected neural networks (FCNNs) that can predict the sea level trend based on climate model projections. Alongside this, our method provides uncertainty estimates associated with the ML prediction. We also show the effectiveness of partitioning our spatial dataset and learning a dedicated ML model for each segmented region. We compare two partitioning strategies: one achieved using domain knowledge and the other employing spectral clustering. Our results demonstrate that segmenting the spatial dataset with spectral clustering improves the ML predictions.

SIGNIFICANCE STATEMENT: Long-term projections are needed to help coastal communities adapt to sea level rise. Forecasting multidecadal sea level change is a complex problem. In this paper, we show the promise of machine learning in producing such forecasts 30 years in advance and over the global ocean. Continued improvements in prediction skills that build on this work will be vital in sea level rise adaptation efforts.

KEYWORDS: Sea level; Climate change; Forecasting; Neural networks; Clustering

1. Introduction

Satellite altimeter observations since 1993 indicate that the global mean sea level is rising at a rate of 3.4 mm yr^{-1} and accelerating by 0.08 mm yr^{-2}, as shown in studies (Nerem et al. 2018; Hamlington et al. 2020a). Global mean sea level has risen 10 cm in the last 30 years. However, there is considerable regional variation in the amount of sea level rise (Hamlington et al. 2016) necessitating the need for a regional sea level change analysis. With three decades of satellite observations, we can now investigate the role played by anthropogenic climate change signals such as greenhouse gasses, aerosols, and biomass burning in this rising sea level. Climate model projections can be used to estimate the extent of the causal contributions from such factors and forecast future sea level changes. In Fasullo and Nerem (2018) and Fasullo et al. (2020b,a), two large ensembles of climate models were studied to show that the forced responses to greenhouse gas and aerosols have begun to emerge in the regional pattern of sea level rise in the altimeter data. This motivates us to utilize climate models in our framework. Our work uses machine learning to predict future regional patterns of sea level change. It is part of a longer-term research project that investigates the extent of contributions from anthropogenic climate-change signals to sea level change. Through our work, we show promising results demonstrating the potential of neural network–based ML models. Our framework uses both satellite observations and climate model simulations to predict sea level trends 30 years into the future at a 2° spatial resolution.

Forecasting long-term sea level change is a complex problem given the natural variability of the ocean, the wide range of processes involved, and the complex nonlinear interactions playing a role in sea level change. Some past studies have used satellite altimeter data and adopted ML techniques to perform sea level prediction. Tide gauge data have also been used for similar tasks, but they suffer from the influence of local coastal effects and poor spatial coverage, while satellite altimeter data provide nearly global coverage and are suitable for working with open ocean sea level patterns. Many tide gauges also suffer from time varying vertical land motion and those that do not have collocated continuous GPS measurements therefore contain significant uncertainty (Watson et al. 2015). Imani et al. (2017) make use of support vector regression for sea level prediction in the tropical Pacific Ocean. In Braakmann-Folgmann et al. (2017), they utilize a combination of convolutional neural network (CNN) + ConvLSTM (Shi et al. 2015) layers to perform interannual sea level anomalies (SLA) prediction over the Pacific Ocean. Zhao et al. (2019) use a combination of least squares and neural networks to produce sea level anomaly prediction in the Yellow Sea. Sun et al. (2020) work with long short-term memory network

Corresponding author: Saumya Sinha, saumya.sinha@colorado.edu

DOI: 10.1175/AIES-D-23-0089.1 e230089

(LSTM) for the South China Sea. Through their work in Liu et al. (2020), the authors employ an attention-based LSTM mechanism for sea surface height (SSH) forecasting in the South China Sea. Balogun and Adebisi (2021) include ocean–atmospheric features like sea surface temperature, salinity, and surface atmospheric pressure to build support vector and LSTM models for the West Peninsular Malaysia coastline. Nieves et al. (2021) make use of Gaussian processes and LSTM to predict sea level variation along the regional coastal zones. In Hassan et al. (2021), they compare various machine learning techniques to predict global mean sea level rise. An important part of the pipeline in Wang et al. (2022) includes a ConvLSTM pipeline consisting of 3D convolutions and attention modules for forecasting altimeter SLA on the South China Sea. These techniques, however, are trained only on the altimeter dataset which to date is only 30 years in length; this can affect the performance of such data-driven models as brought up in this latest survey by Bahari et al. (2023). These approaches also do not use the insights provided by climate model projections that can potentially inform on contributions of anthropogenic climate-change signals. Moreover, these models address regional forecasting with a lead time of a few days to a few years ahead but do not go so far as to forecast sea level change over the global ocean 30 years in advance. Our work utilizes the climate model projections and addresses the problem at a much bigger spatial scale that includes all the oceans and a much longer time horizon in the future. It should be noted that the focus of this work is to predict the sterodynamic component of the sea level (with the global mean removed). The sterodynamic component is the change in the sea level due to changing ocean currents, temperature, and salinity (Gregory et al. 2019).

We work with 30-yr linear trends of the sea level time series. We note that the climate models do not accurately reproduce all aspects of the trend pattern in altimeter data. There is also more variability in the altimeter trends compared to the climate models. We observed this in our previous work (Sinha et al. 2022), where a U-Net (Ronneberger et al. 2015) model is trained on long periods of climate model simulations to produce spatiotemporal predictions 30 years ahead. This U-Net model is then used to predict the future altimeter data. However, these predictions had much lower variability as compared to the altimeter observations. This underscores the challenge of combining modeled and observed fields in producing sea level predictions.

Working with multidecadal global trends severely limits the ground-truth data we have. Thus, we use the sea level trend values at every spatial grid point to create a training dataset for our ML model. With a 2° spatial resolution, we get a 180 × 90 (longitude × latitude) grid in our sea level trend maps. This gives us a reasonably large dataset for training an ML model even for a single 30-yr-long trend for each grid point. We build a supervised learning framework using fully connected neural networks (FCNNs) that learns a nonlinear mapping of the climate model trends to predict the altimeter trend while absorbing the biases that the climate models have away from the altimeter observations. This is accompanied by an interpretability study that explains the contributions of all the climate models to our final prediction. Given that the dominant factors driving sea level variability differ by region, we segment our spatial dataset and learn separate FCNNs for each segmented region. We compare a partition achieved using domain knowledge to a partition achieved via spectral clustering. We show that segmenting the spatial dataset improves the ML predictions. Spectral clustering shows promise by predicting future trends with ML such that their variability lies in the range we expect, given the variability of the past altimeter observations. Our predictions with spectral clustering also have lower uncertainties in impactful areas.

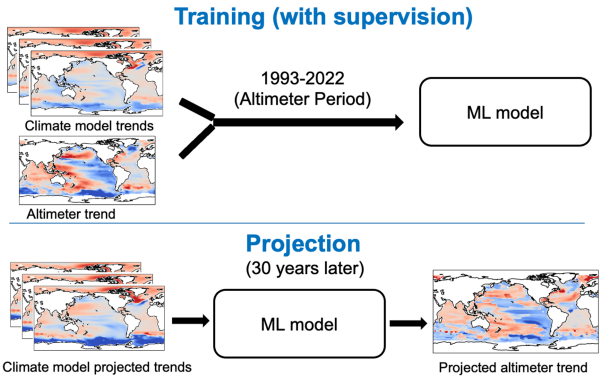

FIG. 1. Overall ML pipeline for the task of sea level trend prediction using trends from climate models projections and altimeter observations.

2. Method

Our supervised learning pipeline is trained for the period 1993–2022. The spatial grid is flattened to create our dataset, where each data point corresponds to an oceanic grid point. For every grid point, a linear trend is computed over 1993–2022 for the climate model ensemble means (described in section 3), comprising the input features X. The trends computed for the altimeter data (ground truth) make the label Y for our supervised ML training. The global mean is removed from both the climate model trends and the altimeter trend. We get the supervision from the altimeter trend Y and the features X to our ML model from the climate model hindcast trends. In the inference phase, we predict trends for 30 years later. This is done by taking the climate model projected trends for 2023–52 and passing them through the learned ML model to predict the altimeter trend. See the overall ML framework in Fig. 1. The ML model is a FCNN trained with mean squared error (MSE) as the loss. We ran experiments with a random forest–based ML model, but it performed poorly when compared to FCNN, showing the latter is more suitable for learning the nonlinear mapping in our case. The MSE is weighted, where the weights are the cosine of the latitude of the grid points. This gives spatial weighting which essentially assigns more weight to the equatorial regions and less weight to polar regions.

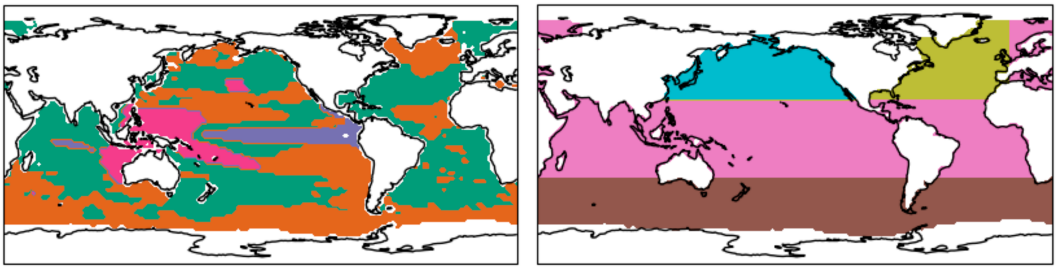

FIG. 2. Spatial segmentations obtained from (a) spectral clustering and (b) a domain-specified partition derived from our physical understanding of the data, where the North Atlantic Ocean (olive) and North Pacific Ocean (cyan) are assigned individual partitions, latitudes from south up to −30 are assigned another partition (brown), and the rest belong to the fourth partition (pink).

a. Clustering

We segment our spatial grid into partitions or clusters and observe the performance of the ML model when trained based on these clusters, i.e., a separate FCNN is trained for each cluster. This is based on the hypothesis that learning ML model weights that are attuned to each cluster can be more optimal than a single ML model for the entire globe. Our study compares spectral clustering against a domain-specified partitioning that is derived from our physical knowledge of the data and proposed by the domain experts in the team. The time series of the altimeter sea surface height (with the seasonality removed) serves as the features for spectral clustering. Empirical evaluation with k-means clustering failed to perform close to spectral clustering and is not included in the study. The spatial segmentations with spectral clustering as well as the domain-specified partition can be seen in Fig. 2. The spectral clustering, as observed by domain experts, seemsto be influenced by the El Niño–Southern Oscillation (ENSO) phenomenon in the Pacific region. This could be because of the similarity between spectral decomposition and empirical orthogonal function (EOF) analysis and the fact that ENSO is the leading mode of interannual climate variability (Vestergaard et al. 2010). This could be beneficial as creating these clusters helps to treat ENSO-specific regions separately. These partitioning strategies are compared to each other and to a setup where the spatial grid is not segmented at all.

b. Hyperparameter tuning and model architecture

We make use of k-fold cross validation ($k = 5$) to choose the best hyperparameters for each cluster, ending up with different FCNN architectures per cluster. To elaborate further, each of the orange and green clusters in the spectral clustering setup as seen in Fig. 2a learns an FCNN consisting of three hidden layers with 1024, 512, and 256 neurons, respectively. For each of the other two smaller clusters, we use an FCNN with two hidden layers and 256 and 128 neurons, respectively. Each hidden layer is followed by a ReLU activation. The $l2(0.000\,005)$ regularizer and a single dropout layer (0.2) are applied to avoid overfitting in each of the ML models.

3. Dataset

Two types of data are used in this study: altimeter data and climate model large ensemble (LE) experiments. The altimeter dataset is a monthly SSH data at 1/4° spatial resolution for the time period 1993–2022. For the same duration, we obtain monthly SSH at 1° spatial resolution from the ensemble means of six different climate model LEs produced with CESM1 (Kay et al. 2015), CESM2 (Danabasoglu et al. 2020), GFDLESM2M (Dunne et al. 2013), MPIGE (Maher et al. 2019), MPI-ESM1-2-HR (Müller et al. 2018), and MPI-ESM1-2-LR (Giorgetta et al. 2013). These LEs provide simulations for the twentieth and twenty first centuries and are multimember ensembles of climate models running with small perturbations in the initial conditions to estimate the distribution of internal climate variability and forced climate change.

a. Preprocessing

Model simulations for individual members of the above large ensembles are averaged to create the SSH variable. We do this since we expect internal variability to be inherently unpredictable while we expect the response to external forcings (influences considered external to the climate system that impact climate) to be both predictable and slowly varying. This step not only removes noise but also reduces variability in the climate model SSH data, while we have variability present in the altimeter SSH data as it is a single field. Some of the climate models operate under assumptions in which their global mean is by definition 0. We, therefore, remove the global mean in all the datasets including the altimeter. A spatialsmoothing[1] is applied on the altimeter field to reduce the influence of small-scale ocean eddies. The spatial SSH fields for both the altimeter data and climate model output are regridded to a 2°, i.e., a 180 × 90 grid as it speeds up the computation while still keeping a reasonable resolution. For every ocean grid point, a linear trend is fitted to the monthly SSH time series for the 1993–2022 (30-yr) time period. This way, a single trend map is obtained, for all the ensembles and the

[1] https://www.ncl.ucar.edu/Document/Functions/Built-in/exp_tapersh.shtml.

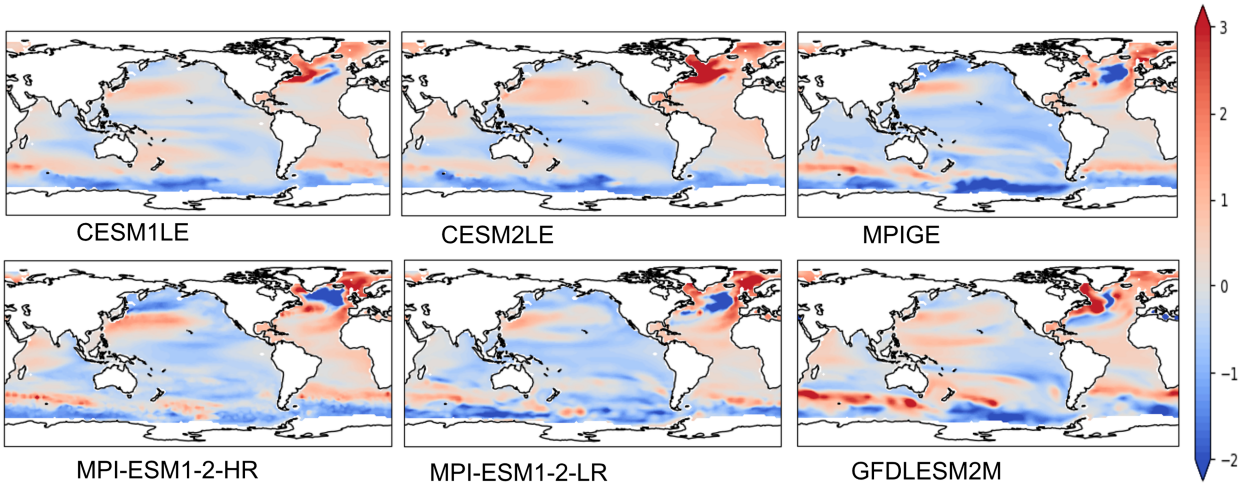

FIG. 3. The sea level trend maps for the six climate model LEs for the period 1993–2022 with their global mean removed. Here, the trend values are visualized in millimeter per year.

altimeter (see Figs. 3 and 4). Working with trends helps to avoid the monthly variability of the SSH fields. We can observe the differences between the altimeter and the climate model trends, especially with respect to variability. The climate models do not accurately reproduce the trend pattern in altimeter data, and there is a lot more variability in the altimeter trend as compared to the climate model trends.

b. *Feature inputs and label for the ML model*

The trends obtained via the above preprocessing are used for the ML training. The climate model trends act as the input X to our ML model, whereas the altimeter trends act as the ground truth labels Y for our ML model.

It is worth noting that altimeter records are not present for all latitudes. We have both altimeter and climate model trend values for 8001 global ocean points (excluding land grid points) that we use as the dataset for ML. We show trend values from the six climate models in Fig. 3 that serve as the input features X for our ML model and the altimeter trend value in Fig. 4 that serves as the label Y for training the ML model. To summarize, the final dataset prepared for ML is a tabular one, where the rows are the ocean grid points, and for each row, we have the feature inputs given by the climate model trends and label given by the altimeter trend. These trend values are computed in centimeter per year and are normalized by scaling them between 0 and 1 for training. After training, in the inference phase, trends are predicted for 30 years later. Climate model projected trends are computed in the same way for 30 years later, i.e., for 2023–52. These are then passed through the learned ML model to predict the altimeter trend for 2023–52.

To reiterate, the focus of this work is to predict the sterodynamic component of the sea level trend—with the global mean removed.

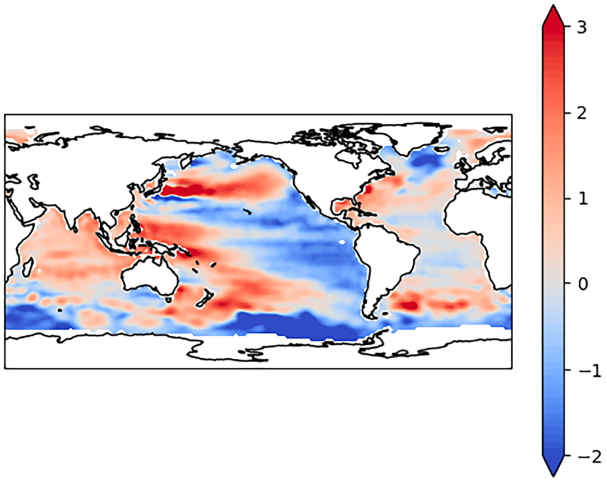

FIG. 4. The altimeter sea level trend map for the period 1993–2022 with the global mean removed. Here, the trend values are visualized in millimeter per year.

4. Results

We report our results using different evaluation metrics for the past and future time periods, since there is no ground truth with which to evaluate future predictions.

a. *1993–2022*

With the ground truth data available for this period, in Table 1, we report the RMSE and mean absolute error (MAE)

TABLE 1. Comparing the ML prediction performance in terms of weighted RMSE (mm yr^{-1}), MAE (mm yr^{-1}), and correlation for different spatial segmentations for 1993–2022.

Method	RMSE↓	MAE↓	Correlation↑
No clustering	0.72	0.51	0.82
Domain-specified partition	0.4	0.27	0.95
Spectral clustering	0.51	0.36	0.91

scores on the historical (training) time period, for the two spatial segmentation strategies, compared to applying our supervised learning step directly to the entire spatial extent (no clustering). Table 1 also shows the Pearson correlation scores between the ML predicted trend and the true altimeter trend for 1993–2022. The RMSE, MAE, and correlation scores are spatially weighted as described in section 2. The domain-specified partition is observed to have better scores (lower RMSE, MAE, and higher correlation) for the training period as compared to spectral clustering. Both the domain-specified partition and spectral clustering scores are considerably better than the no clustering setup. Each of the segmented regions is examined by looking at each cluster's RMSE, MAE, and correlation scores. The trend predicted by ML is visualized, and higher error zones are mostly observed in the green cluster of Fig. 2a for spectral clustering and the olive cluster of Fig. 2b for domain-specified partition.

While these scores explain the ML's training performance, our interest mainly lies in the future period prediction which is detailed below.

b. 2023–52

It is harder to gauge the performance of any ML method without the ground truth. In this case, we do a qualitative analysis of the predicted trend in terms of cumulative variability, to evaluate the ability of the ML models to predict trends with variability similar to the variability of the 1993–2022 altimeter trend. Additionally, we compute the model uncertainty of the ML models in their prediction. As often done in the climate science domain, we also evaluate the ML models solely with the climate model datasets (Monteleoni et al. 2011). These experiments are described later in this section.

We use the root-mean-square (RMS) value of the trend (spatially weighted as in section 2) to quantify the notion of variability in the trend. The RMS value is higher if the cumulative variability is higher and vice versa. Figure 4 shows that the altimeter trend from 1993 to 2022 has a high variability. We computed the RMS value to be 1.23 mm yr^{-1}. This gives us the baseline variability of persistence, a standard baseline approach in climate and weather forecasting, i.e., considering this observed variability as an estimate of future variability. In Figs. 5a–c), we show the future predicted trend obtained from the ML model without any partitioning (no clustering), with the domain-specified partition, and with a partition obtained via spectral clustering, respectively. We computed the RMS values associated with trend predictions obtained from the three strategies. Trend predicted with spectral clustering (Fig. 5c) shows a high variability with RMS as 1.05 mm yr^{-1} for 2023–52. It is very close, though still slightly less than the altimeter trend variability of the past. On the other hand, the trend predicted with the domain-specified partition (Fig. 5b) shows a much higher variability with RMS as 1.68 mm yr^{-1}. The high prediction red zone in the North Pacific Ocean could be dominating the overall RMS value of the domain-specified prediction. This emphasizes the need to use additional metrics and analyses to evaluate our predictions rather than relying on a single overall

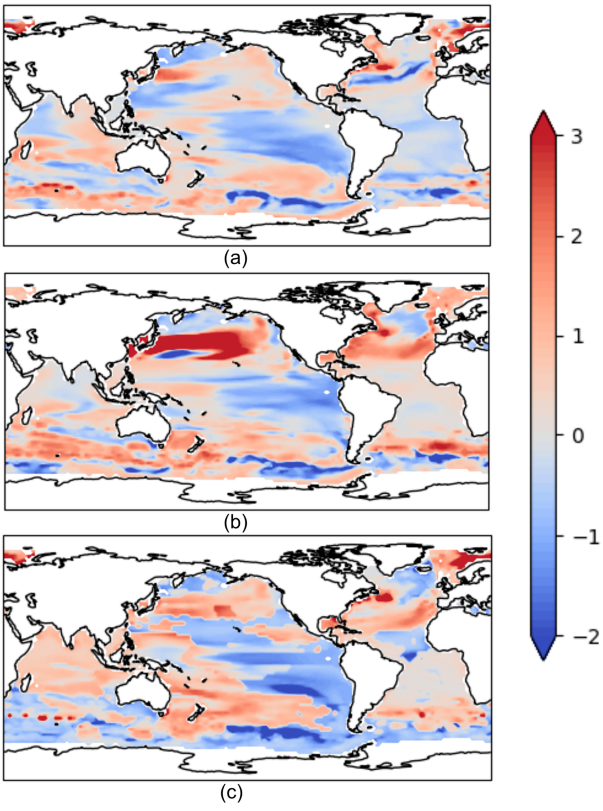

FIG. 5. The trend estimates are predictions for the future period: 2023–52 in millimeter per year. The trend predicted with ML using (a) no partitioning of the spatial grid (RMS: 0.81 mm yr^{-1}), (b) the domain-specified partition (RMS: 1.68 mm yr^{-1}), and (c) spectral clustering (RMS: 1.05 mm yr^{-1}).

score expressing cumulative variability. Notably, the predicted variability of both spectral clustering and the domain-specified partition is higher as compared to the no clustering setting (RMS: 0.81 mm yr^{-1}). This result strengthens our hypothesis that segmenting the spatial grid and learning one ML model on each segmented region yield predictions that can better capture variability (with respect to persistence). Measuring the correlation between the persistence and future predicted trend is also useful as it is expected to be fairly high based on the climate model experiments which also show a high correlation between the past and future trend in their projections. This correlation is much higher (0.59) for spectral clustering than the domain-specified partition (0.45) and no clustering setup (0.27).

1) MODEL UNCERTAINTY

Providing uncertainties of machine learning predictions can be extremely useful. For this application, we do so as another way to evaluate our ML model's future predictions. Gal and Ghahramani (2016) showed theoretically that neural networks with dropout layers can be interpreted as a Bayesian approximation of a deep Gaussian process. Thus, we can obtain uncertainties with dropout neural networks without sacrificing accuracy and with lesser computation cost as compared to the

Bayesian models. This Monte Carlo dropout approach can work with any existing neural networks trained with dropout (Gal and Ghahramani 2016). This essentially simulates having multiple models and helps to assess our ML model's robustness in terms of its prediction uncertainty.

Our FCNN model includes dropout layers to reduce overfitting while training, thus allowing us to use the Monte Carlo dropout approach for uncertainty estimation. To do so, in the inference phase, we perform multiple forward passes (with different dropout masks) through our ML model. We then report the mean of the ensemble of predictions as the prediction outcome and their standard deviation as the prediction uncertainty. Figure 6 shows the prediction uncertainty plots for both spectral clustering and the domain-specified partition. The predictions with the domain-specified partition (Fig. 6a) show a higher overall variance in prediction. We observe higher uncertainties in key areas that are critical for socioeconomic impacts such as important parts of the Pacific Ocean, whereas spectral clustering (Fig. 6b) predictions are more confident in most of the Pacific Ocean and higher uncertainties are concentrated in the Southern Ocean and parts of the North Atlantic Ocean. We also studied the cumulative uncertainty by taking the RMS of this model uncertainty over the global ocean. Lower RMS is better as it indicates lower cumulative uncertainty. The RMS for spectral clustering (0.19 mm yr^{-1}) is better than the domain-specified partition (0.24 mm yr^{-1}) and the no clustering scenario (0.3 mm yr^{-1}). We observe that the ML model is more certain with spectral clustering.

We also observed that some regions over which the ML model with spectral clustering had higher uncertainties (Fig. 6b) had high overlap with the regions where climate model projections for 2023–52 had the highest disagreement (Fig. 8b).

2) Interpretability study

Through this interpretability study, our goal was to understand the contribution of each climate model in the ML prediction. While complex machine learning models can predict accurate outcomes, it is extremely important to understand why the ML model makes a certain prediction in order to make it more interpretable. We use shapley additive explanation (SHAP) Lundberg and Lee (2017) to compute the contributions of each feature to a prediction outcome in order to explain the prediction.

Lundberg and Lee (2017) in their work on SHAP show the value of a linear explanation model that is an interpretable approximation to the original complex model by proposing a class of methods: additive feature attribution methods. They use x and f as the original inputs and prediction model, g as an explanation model, and x' as a simplified input such that $x = h_x(x')$, h_x being a mapping function. Under the definition of additive feature attribution methods as described in (Lundberg and Lee 2017), the explanation model g must satisfy $g(z') \approx f[h_x(z')]$, where $z' \approx x'$, and can be written as $g(z') = \phi_0 + \sum_{n=1}^{M} \phi_i z_i'$, $z' \in \{0, 1\}^M$ indicates whether a particular feature (out of M features) is included or not with a binary value. Here, ϕ_i indicates feature attribution or feature importance, i.e., how much this feature contributed to the model's outcome. Lundberg et al.

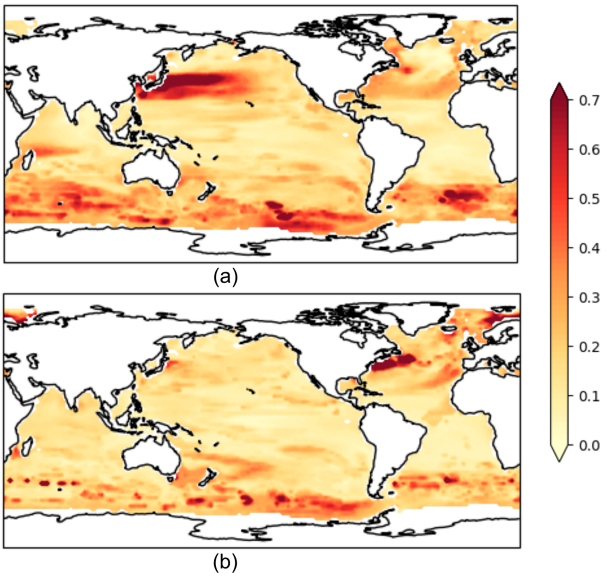

FIG. 6. ML model uncertainty map in terms of standard deviation (mm yr^{-1}) over future prediction with (a) the domain-specified partition and (b) spectral clustering.

leverage the game theory literature to show that Shapley values as ϕ_i satisfy the definition and a few more desirable properties of this class of methods (Shapley 1953; Young 1985). For the computation of Shapley values (Lipovetsky and Conklin 2001), marginal contribution of a feature i is computed by taking the difference between the model f's output with and without that feature. The marginal contribution is computed for all possible subsets $S \subseteq F \setminus \{i\}$ (F is the set of all features), and a weighted average over them gives the Shapley value as shown below (from Lundberg and Lee 2017):

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)]. \quad (1)$$

In most cases, ML models cannot handle missing features, so this is often approximated by integrating out the feature using samples from a background dataset as discussed in Štrumbelj and Kononenko (2014) and Lundberg and Lee (2017). The computation of SHAP becomes very challenging as the number of features increases. In Lundberg and Lee (2017), they provide an approximation to obtaining the Shapley values via Kernel SHAP (a model agnostic approximation).

Our interpretability analysis is based on SHAP as explained above. We use Kernel SHAP from Python's shap.KernelExplainer to compute the contributions or feature importance values of the climate models which are feature inputs to our FCNNs in order to explain the future prediction. We apply SHAP on each of the clusters since we learn different ML models for different clusters. Figure 7 shows a cluster level feature importance ranking of all the climate models for the spectral clustering setup. SHAP assigns a feature importance to each climate model for the future prediction on each grid point in the cluster. These importance values are averaged over every cluster and shown in the bar plots in Fig. 7. Given

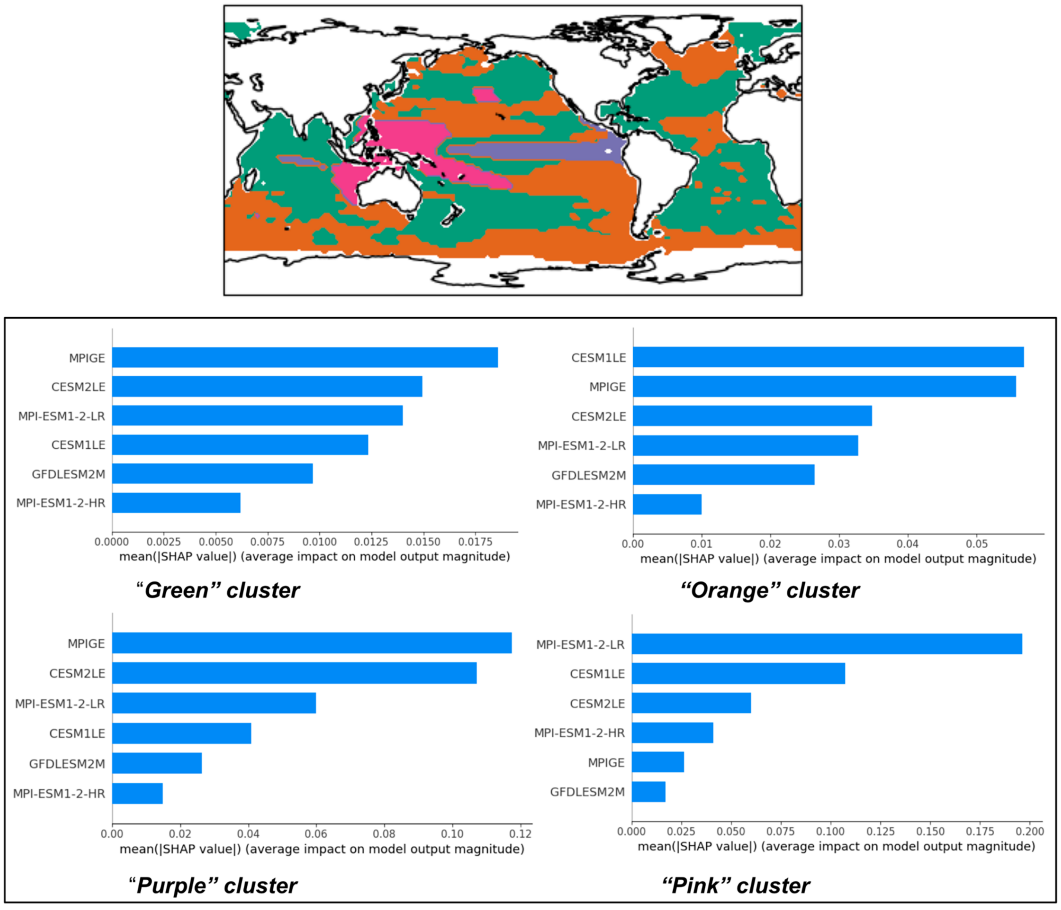

FIG. 7. (top) The clusters obtained from spectral clustering and (bottom) the plot showing the feature importance ranking of all the climate models based on their contribution to the future prediction as given by SHAP values. The map with the spectral clusters is added to provide more context.

that we use scaling on the input features as well as the output labels when training the ML models, the importance values displayed in Fig. 7 are scaled as well. Overall, the SHAP values indicate that the CESM1, CESM2, and Max Planck Institute Grand Ensemble (MPIGE) large ensembles are more important for all the clusters than others, suggesting that ML model relied more on these for future predictions.

Furthermore, in order to verify the consistency of the feature rankings, we used another popular explainability method: local interpretable model-agnostic explanations (LIME) (Ribeiro et al. 2016), as an additional method to compute the feature contributions and provide a ranking of the features in our ML model. Overall, we note a consistency in the feature rankings generated by LIME and SHAP. We see a strong match in feature rankings obtained via LIME when the features had distinctly high or low scores with SHAP, suggesting high consistency when features have distinctly high or low scores compared to other features. For example, we see a match in the best feature for the "Green" (top two features match) and "Pink" clusters and a match in the worst features of "Orange" and "Purple" clusters. Both methods yielded almost the same top three features with a slight change in their top three ordering and similarly almost the same bottom three features across all clusters. The slight change of ordering was common where SHAP ranked features were not significantly different.

3) EVALUATION WITH CLIMATE MODELS

While we do not have observations for the future to validate our prediction outcomes, we do have climate model projections for the future. As often done in the climate science domain, we perform an evaluation of our predictions using only the climate model datasets. We train the same FCNN models for 1993–2022 again, but this time using one of the climate model hindcasts as the training label instead of the altimeter data, and the rest of the five climate models as input features (like in Monteleoni et al. 2011). We train six such ML models treating each of the six climate models as the training label one at a time. At the time of inference for 2023–52, we have the ground truth, i.e., climate model projections available for the future for each climate model, so we measure RMSE, MAE, and correlation scores for the ML prediction of the climate model trend against the true climate

TABLE 2. An evaluation setup using climate models to simulate observation data so as to evaluate ML predictions for the future period 2023–52. The performance score in terms of (a) correlation (higher the better) and (b) RMSE and (c) MAE (in mm yr^{-1}, lower the better) between the ML prediction of a climate model and the ground truth climate model projection. Results are computed for predicting each of the six climate models one at a time (shown as the columns in the table) using the rest of them as input features. The last column shows an average score obtained by averaging the scores over all climate models. The best results are highlighted in bold.

	CESM1 LE	CESM2 LE	MPIGE	MPI-ESM1-2-HR	MPI-ESM1-2-LR	GFDL ESM2M	Average
			(a) Correlation↑				
Persistence	0.74	0.74	0.73	**0.53**	0.49	**0.74**	0.66
ML with domain-specified partition	0.79	0.73	0.74	0.37	0.49	0.48	0.6
ML with spectral clustering	**0.82**	**0.81**	**0.82**	0.39	**0.6**	0.43	0.65
			(b) RMSE↓				
Persistence	0.58	0.6	0.69	**0.83**	0.95	**0.72**	0.73
ML with domain-specified partition	0.56	0.69	0.73	0.96	0.99	0.99	0.82
ML with spectral clustering	**0.49**	**0.56**	**0.58**	0.93	**0.86**	1.01	0.74
			(c) MAE↓				
Persistence	0.39	0.41	0.38	**0.46**	0.6	**0.46**	0.45
ML with domain-specified partition	0.38	0.46	0.47	0.56	0.6	0.64	0.52
ML with spectral clustering	**0.33**	**0.35**	**0.37**	0.53	**0.56**	0.64	0.46

model projected trend. We show the weighted correlation metrics in Table 2a, weighted RMSE scores in Table 2b, and weighted MAE scores in Table 2c for both the spectral clustering and domain-specified partition setups. They are evaluated against the persistence scores for each of the climate models (here, persistence is using the climate model hindcast from 1993 to 2022 as the prediction for the future 2023–52). It should be noted that the climate model projections substantially differ from each other which makes this prediction task harder for ML. Figure 8 shows the trend maps from all the climate models for the future period 2023–52 and a standard deviation plot showing the variance in their projections.

Based on the correlation, RMSE, and MAE scores, it can be seen that ML with spectral clustering outperforms the domain-specified partition on nearly all the climate models, falling slightly behind only for the case of GFDLESM2M. It also performs better than the persistence on all the climate models except MPI-ESM1-2-HR and GFDLESM2M. We observed that the regions where MPI-ESM1-2-HR and GFDLESM2M predictions with the spectral clustering setup have higher errors (in parts of the Southern Ocean and the North Atlantic Ocean) are some of the regions where these two climate models have disagreement over with the remaining climate model projections (see Fig. 8). Comparing the average correlation, RMSE and MAE scores (last column in Table 2) over all the six ML models based on the six climate model labels show spectral clustering to be better than the domain-specified partition and very close to the persistence.

4) EXPERIMENT WITH VARYING NUMBER OF CLUSTERS

We do a comparative study by varying the number of clusters (*n clusters*) obtained with spectral clustering and comparing their prediction performance based on the evaluation schemes discussed before. Specifically, for *n clusters* as 2, 4, 8, 16, 32, and 64, we present Table 3 where we compare their training error in terms of RMSE and MAE, cumulative variability of the future trend prediction in terms of its RMS, and the ML model uncertainty in prediction quantified by the RMS of model uncertainty. We also include the correlation of the predicted trend with the past altimeter trend (1993–2022). Additionally, we add another column which provides spectral clustering's performance scores when evaluated solely with the climate models. This last column reports an average correlation score as derived in section 4b(3).

With an increase in the number of clusters, there are fewer data points per cluster, so the training data size for each ML model decreases. Table 3 indicates that the training RMSE and MAE decrease with increasing *n clusters*. This is expected as the training process will tend to overfit more with smaller training data per cluster. The RMS that represents the cumulative variability of the future prediction outcome is observed to increase with the increase in *n clusters* (except for a small drop for *n clusters* = 32). Notably, the model uncertainty drops and then increases, especially when working with a larger number of clusters like *n clusters* = 32 or 64, as quantified by the RMS in the third column. For such high *n clusters*, there is a huge decrease in the training data points per cluster and this can lead to more variance in ML's prediction, reducing its confidence. Higher *n clusters* show predicted trends to be generally more correlated with the past altimeter trend. The last column based on evaluation with climate models does not show a significant performance change with *n clusters*. The score, however, drops slowly with more *n clusters*. For a qualitative comparison, we plot the predicted trend for 2023–52 as generated by the ML model with 4, 8, and 16 spectral clusters in Fig. 9.

While we observe slightly better predictions with eight spectral clusters (from Table 3), we work extensively with the 4-cluster spectral clustering setup in order to have a fair comparison with the domain-specified partition with four partitions in our case. Having a higher number of clusters also makes it harder for the domain experts to interpret its physical implications. Additionally, upon examining the prediction maps closely from Fig. 9, it can be noted that the difference across various clusters can mostly be seen in the predicted strength of the trends (higher for higher *n clusters* which also

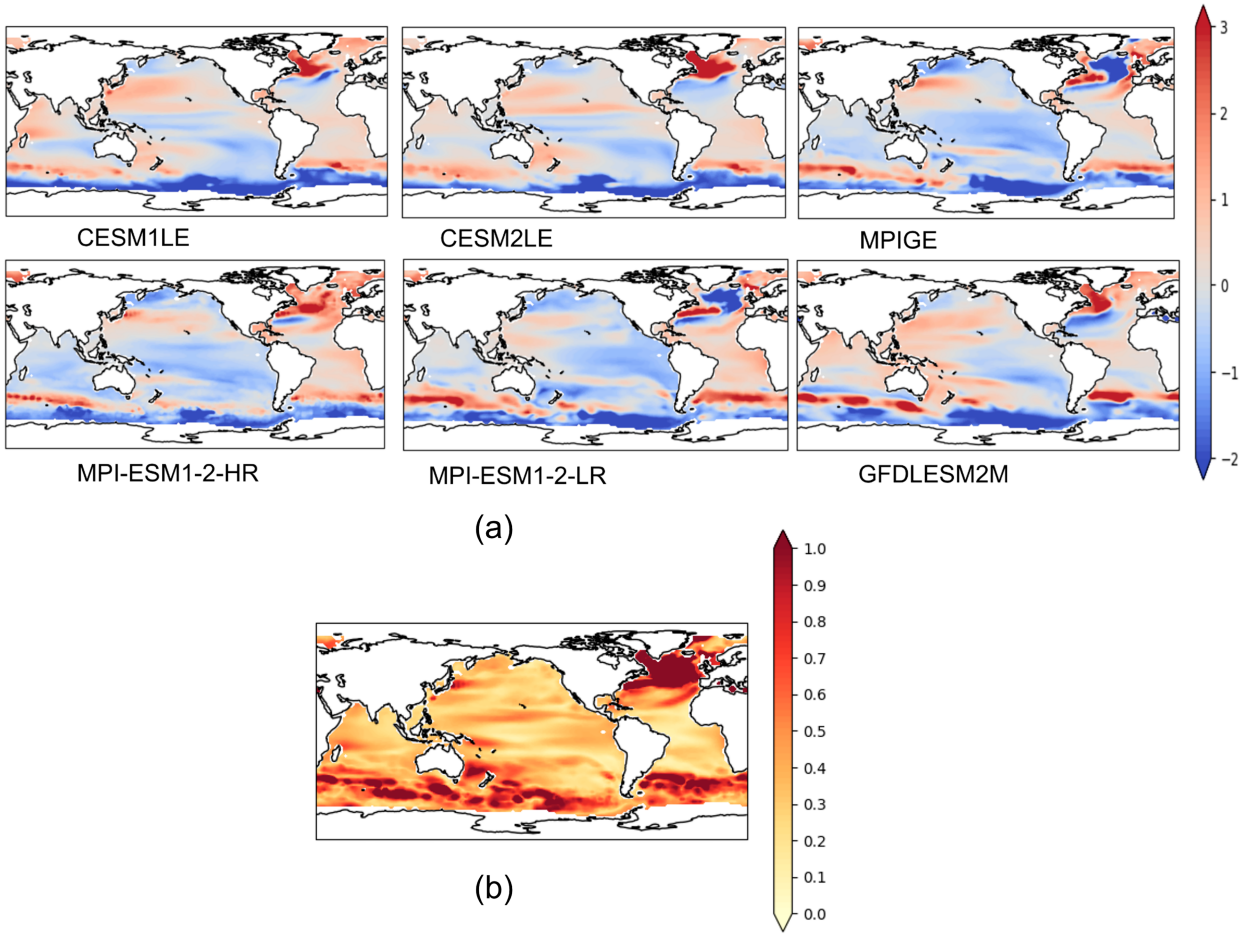

FIG. 8. Plot showing (a) the climate model projected trends in millimeter per year for 2023–52 and (b) the standard deviation in their projections in millimeter per year.

contributes to its higher RMS) and notably not the general prediction patterns themselves.

5) EXPERIMENT WITH MODEL DRIFT REMOVED FROM CLIMATE MODELS

We perform an experiment with the spectral clustering setup (with four clusters) where we remove the model drift from the climate models and use them in our ML pipeline to obtain the future trend predictions. The drift is computed as the linear trend from the overlapping 250 years from the preindustrial control run using annual mean at each latitude and longitude and subtracting that drift from the historical and future estimates. Figure 10 shows the trend prediction for 2023–52 as a result of this experiment. We observe the RMS to be

TABLE 3. Comparing ML with spectral clustering performance across *n clusters*: 2, 4, 8, 16, 32, and 64. It shows the training errors (RMSE and MAE), RMS of the future predicted trend, RMS of the ML model uncertainty in prediction, correlation of the future predicted trend with the past altimeter trend, and an average correlation score when evaluated only with the climate model datasets as described in section 4b(3). All the scores are weighted and the RMSE, MAE, and RMS measures are in millimeter per year.

n clusters	Training RMSE	Training MAE	RMS of future predicted trend (cumulative variability	RMS of ML model uncertainty	Correlation of predicted trend with past altimeter trend	Avg correlation on evaluation with climate models
2	0.62	0.44	0.99	0.23	0.45	0.68
4	0.51	0.36	1.05	0.19	0.59	0.65
8	0.43	0.29	1.11	0.17	0.69	0.67
16	0.34	0.23	1.5	0.17	0.67	0.62
32	0.38	0.26	1.29	0.24	0.69	0.62
64	0.29	0.2	1.61	0.25	0.67	0.64

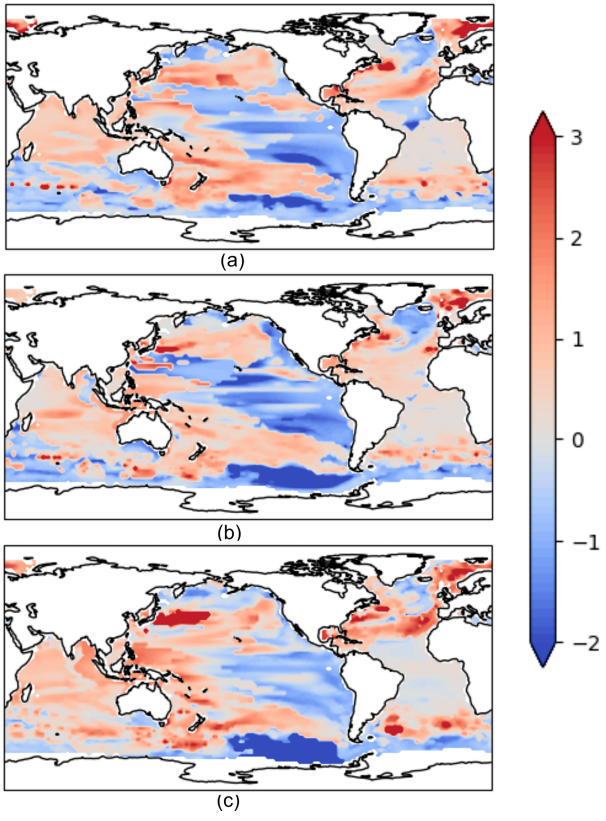

FIG. 9. The ML predicted trend in millimeter per year for the future period 2023–52 with spectral clustering with (a) 4, (b) 8, and (c) 16 clusters.

1.23 mm yr^{-1} and its correlation with the past altimeter trend to be 0.65, which are slightly higher than our original spectral clustering experiment setup with four clusters (see Fig. 5c). These results without the model drift are similar and sandwiched between the results we see with the varying number of clusters without the model drift removed (see Table 3). Further investigation into this will be conducted in future studies.

5. Discussion

In our framework, fully connected neural networks learn to map climate model projections to altimeter trends. We also present an interpretability study that uses SHAP values to explain the contributions of all the climate models to the final prediction. Spectral clustering shows promise in this application by generating future predictions with ML such that their variability lies in the range we expect, given the variability of the past altimeter observations. These ML predictions have lower uncertainties in impactful areas as shown before. Spectral clustering also shows robustness as it yields better predictions than the one with the domain-specified partition when evaluated solely with climate models, as described in section 4b(3). The future predictions are expected to be correlated well with the past altimeter trend, and a higher correlation is observed with the predictions obtained from spectral clustering than the domain-specified partition.

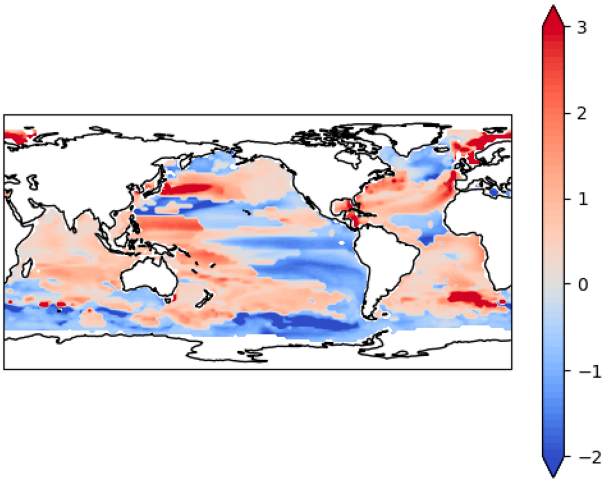

FIG. 10. The ML predicted trend in millimeter per year for the future period 2023–52 using climate models after removing the model drift in them (spectral clustering with four clusters setup).

Our prediction highlights the regional variation in the predicted sea level trend. It is worth noting that the climate model projections used in our framework are "RCP85" and "SSP370" scenarios. These mostly lie in the high level of emission scenarios where a few policies have been put in place to reduce emissions and warming and tackle climate change. Under such circumstances, our prediction outcomes indicate a rising sea level trend, without considering the global mean sea level (GMSL) change, around regions such as Japan, India, the South China Sea, the Maritime Continent, Australia, the Gulf Coast and the eastern seaboard of the United States, and Mexico. Overall, these predictions suggest that many existing hotspots of sea level rise, including highly populated zones in the western Pacific Ocean and along the U.S. Gulf Coast, will continue to experience rates of sea level rise in excess of the global average (GMSL). Some of these areas may be limited in their ability to adapt to such changes which could increase the risk of major impacts of sea level rise in the coming decades.

6. Conclusions

We show the effectiveness of neural networks in multidecadal sea level trend prediction at a 2° spatial grid leveraging the projections from climate model large ensembles. We demonstrate that segmenting the spatial grid into partitions employing spectral clustering improves the ML predictions by learning a dedicated ML model per partition. We also supplement our predictions with uncertainty estimates which could be more helpful in interpreting the results. While our framework presents promising results, it is important to note that climate model projections become less certain over time, making long-term predictions based on them challenging. The climate models used in our setup do not incorporate melting ice sheets and their effects on future sea level change (Hamlington et al. 2020b). It is pertinent to utilize this to improve the predictions further. The predictions can potentially

also improve if we incorporate factors such as wind and temperature, harnessing deep neural networks' capabilities in handling these diverse data.

Acknowledgments. This work was supported by NASA Award 80NSSC21K1191. The efforts of CM in this work were supported by NSF Cooperative Agreement 2153040. The author thanks Shivendra Agrawal for their valuable feedback on improving the quality of this work.

Data availability statement. The climate model output used in this study can be accessed on the Earth System Grid at https://esgf-node.llnl.gov/search/cmip6/. The sea level altimeter data from JPL can be accessed at https://sealevel.nasa.gov/data/dataset/?identifier=SLCP_SEA_SURFACE_HEIGHT_ALT_GRIDS_L4_2SATS_5DAY_6THDEG_V_JPL2205_2205.

REFERENCES

Bahari, N. A. A. B. S., A. N. Ahmed, K. L. Chong, V. Lai, Y. F. Huang, C. H. Koo, J. L. Ng, and A. El-Shafie, 2023: Predicting sea level rise using artificial intelligence: A review. *Arch. Comput. Methods Eng.*, **30**, 4045–4062, https://doi.org/10.1007/s11831-023-09934-9.

Balogun, A.-L., and N. Adebisi, 2021: Sea level prediction using ARIMA, SVR and LSTM neural network: Assessing the impact of ensemble ocean-atmospheric processes on models' accuracy. *Geomatics Nat. Hazards Risk*, **12**, 653–674, https://doi.org/10.1080/19475705.2021.1887372.

Braakmann-Folgmann, A., R. Roscher, S. Wenzel, B. Uebbing, and J. Kusche, 2017: Sea level anomaly prediction using recurrent neural networks. arXiv, 1710.07099v1, https://doi.org/10.48550/arXiv.1710.07099.

Danabasoglu, G., and Coauthors, 2020: The Community Earth System Model version 2 (CESM2). *J. Adv. Model. Earth Syst.*, **12**, e2019MS001916, https://doi.org/10.1029/2019MS001916.

Dunne, J. P., and Coauthors, 2013: GFDL's ESM2 global coupled climate–carbon Earth System Models. Part II: Carbon system formulation and baseline simulation characteristics. *J. Climate*, **26**, 2247–2267, https://doi.org/10.1175/JCLI-D-12-00150.1.

Fasullo, J. T., and R. S. Nerem, 2018: Altimeter-era emergence of the patterns of forced sea-level rise in climate models and implications for the future. *Proc. Natl. Acad. Sci. USA*, **115**, 12 944–12 949, https://doi.org/10.1073/pnas.1813233115.

——, P. R. Gent, and R. Nerem, 2020a: Forced patterns of sea level rise in the Community Earth System Model Large Ensemble from 1920 to 2100. *J. Geophys. Res. Oceans*, **125**, e2019JC016030, https://doi.org/10.1029/2019JC016030.

——, ——, and R. S. Nerem, 2020b: Sea level rise in the CESM Large Ensemble: The role of individual climate forcings and consequences for the coming decades. *J. Climate*, **33**, 6911–6927, https://doi.org/10.1175/JCLI-D-19-1001.1.

Gal, Y., and Z. Ghahramani, 2016: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proc. 33rd Int. Conf. on Machine Learning*, New York, NY, PMLR, 1050–1059, https://proceedings.mlr.press/v48/gal16.html.

Giorgetta, M. A., and Coauthors, 2013: Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5. *J. Adv. Model. Earth Syst.*, **5**, 572–597, https://doi.org/10.1002/jame.20038.

Gregory, J. M., and Coauthors, 2019: Concepts and terminology for sea level: Mean, variability and change, both local and global. *Surv. Geophys.*, **40**, 1251–1289, https://doi.org/10.1007/s10712-019-09525-z.

Hamlington, B. D., S. H. Cheon, P. R. Thompson, M. A. Merrifield, R. S. Nerem, R. R. Leben, and K.-Y. Kim, 2016: An ongoing shift in Pacific Ocean sea level. *J. Geophys. Res. Oceans*, **121**, 5084–5097, https://doi.org/10.1002/2016JC011815.

——, C. G. Piecuch, J. T. Reager, H. Chandanpurkar, T. Frederikse, R. S. Nerem, J. T. Fasullo, and S.-H. Cheon, 2020a: Origin of interannual variability in global mean sea level. *Proc. Natl. Acad. Sci. USA*, **117**, 13 983–13 990, https://doi.org/10.1073/pnas.1922190117.

——, and Coauthors, 2020b: Understanding of contemporary regional sea-level change and the implications for the future. *Rev. Geophys.*, **58**, e2019RG000672, https://doi.org/10.1029/2019RG000672.

Hassan, K. M. A., M. A. Haque, and S. Ahmed, 2021: Comparative study of forecasting global mean sea level rising using machine learning. *2021 Int. Conf. on Electronics, Communications and Information Technology (ICECIT)*, Khulna, Bangladesh, Institute of Electrical and Electronics Engineers, 1–4, https://doi.org/10.1109/ICECIT54077.2021.9641339.

Imani, M., Y.-C. Chen, R.-J. You, W.-H. Lan, C.-Y. Kuo, J.-C. Chang, and A. Rateb, 2017: Spatiotemporal prediction of satellite altimetry sea level anomalies in the tropical Pacific Ocean. *IEEE Geosci. Remote Sens. Lett.*, **14**, 1126–1130, https://doi.org/10.1109/LGRS.2017.2699668.

Kay, J. E., and Coauthors, 2015: The Community Earth System Model (CESM) Large Ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bull. Amer. Meteor. Soc.*, **96**, 1333–1349, https://doi.org/10.1175/BAMS-D-13-00255.1.

Lipovetsky, S., and M. Conklin, 2001: Analysis of regression in game theory approach. *Appl. Stochastic Models Bus. Ind.*, **17**, 319–330, https://doi.org/10.1002/asmb.446.

Liu, J., B. Jin, L. Wang, and L. Xu, 2020: Sea surface height prediction with deep learning based on attention mechanism. *IEEE Geosci. Remote Sens. Lett.*, **19**, 1–5, https://doi.org/10.1109/LGRS.2020.3039062.

Lundberg, S. M., and S.-I. Lee, 2017: A unified approach to interpreting model predictions. *Proc. 31st Int. Conf. Neural Information Processing Systems*, Long Beach, CA, Curran Associates Inc., 4768–4777, https://dl.acm.org/doi/10.5555/3295222.3295230.

Maher, N., and Coauthors, 2019: The Max Planck Institute Grand Ensemble: Enabling the exploration of climate system variability. *J. Adv. Model. Earth Syst.*, **11**, 2050–2069, https://doi.org/10.1029/2019MS001639.

Monteleoni, C., G. A. Schmidt, S. Saroha, and E. Asplund, 2011: Tracking climate models. *Stat. Anal. Data Min.*, **4**, 372–392, https://doi.org/10.1002/sam.10126.

Müller, W. A., and Coauthors, 2018: A higher-resolution version of the Max Planck Institute Earth System Model (MPI-ESM1.2-HR). *J. Adv. Model. Earth Syst.*, **10**, 1383–1413, https://doi.org/10.1029/2017MS001217.

Nerem, R. S., B. D. Beckley, J. T. Fasullo, B. D. Hamlington, D. Masters, and G. T. Mitchum, 2018: Climate-change–driven accelerated sea-level rise detected in the altimeter era. *Proc. Natl. Acad. Sci. USA*, **115**, 2022–2025, https://doi.org/10.1073/pnas.1717312115.

Nieves, V., C. Radin, and G. Camps-Valls, 2021: Predicting regional coastal sea level changes with machine learning. *Sci. Rep.*, **11**, 7650, https://doi.org/10.1038/s41598-021-87460-z.

Ribeiro, M. T., S. Singh, and C. Guestrin, 2016: "Why should I trust you?": Explaining the predictions of any classifier. *KDD'16: The 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Francisco, CA, Association for Computing Machinery, 1135–1144, https://doi.org/10.1145/2939672.2939778.

Ronneberger, O., P. Fischer, and T. Brox, 2015: U-Net: Convolutional networks for biomedical image segmentation. *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, Munich, Germany, Springer, 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.

Shapley, L. S., 1953: A value for n-person games. *Contributions to the Theory of Games II*, H. W. Kuhn and A. W. Tucker, Eds., Princeton University Press, 307–318.

Shi, X., Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, 2015: Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *NIPS'15: Proc. 28th Int. Conf. on Neural Information Processing Systems – Vol. 1*, Montreal, Quebec, Canada, MIT Press, 802–810, https://dl.acm.org/doi/10.5555/2969239.2969329.

Sinha, S., C. Monteleoni, J. Fasullo, and R. S. Nerem, 2022: Sea-level projections via spatiotemporal deep learning from altimetry and CESM large ensembles. *2022 Fall Meeting*, Chicago, IL, Amer. Geophys. Union, Abstract OS36A-05.

Štrumbelj, E., and I. Kononenko, 2014: Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.*, **41**, 647–665, https://doi.org/10.1007/s10115-013-0679-x.

Sun, Q., J. Wan, and S. Liu, 2020: Estimation of sea level variability in the China Sea and its vicinity using the Sarima and LSTM models. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **13**, 3317–3326, https://doi.org/10.1109/JSTARS.2020.2997817.

Vestergaard, J. S., A. A. Nielsen, and O. B. Andersen, 2010: Seventeen years of global SSH anomalies analyzed by a maximum information based extension to EOF analysis. 2 pp., https://backend.orbit.dtu.dk/ws/portalfiles/portal/51205191/PortlandPoster.pdf.

Wang, G., X. Wang, X. Wu, K. Liu, Y. Qi, C. Sun, and H. Fu, 2022: A hybrid multivariate deep learning network for multistep ahead sea level anomaly forecasting. *J. Atmos. Oceanic Technol.*, **39**, 285–301, https://doi.org/10.1175/JTECH-D-21-0043.1.

Watson, C. S., N. J. White, J. A. Church, M. A. King, R. J. Burgette, and B. Legresy, 2015: Unabated global mean sea-level rise over the satellite altimeter era. *Nat. Climate Change*, **5**, 565–568, https://doi.org/10.1038/nclimate2635.

Young, H. P., 1985: Monotonic solutions of cooperative games. *Int. J. Game Theory*, **14**, 65–72, https://doi.org/10.1007/BF01769885.

Zhao, J., Y. Fan, and Y. Mu, 2019: Sea level prediction in the yellow sea from satellite altimetry with a combined least squares-neural network approach. *Mar. Geod.*, **42**, 344–366, https://doi.org/10.1080/01490419.2019.1626306.