JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2022MS003426

Key Points:

- Machine learning tools are used to improve model predictive skill using observationally based data sets
- Our method has great potential to extend the current prediction length from about 7 days to 11–15 days without significantly compromising accuracy

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

G. Liu,
gliu87@hawaii.edu

Citation:

Liu, G., Bracco, A., & Brajard, J. (2023). Systematic bias correction in ocean mesoscale forecasting using machine learning. *Journal of Advances in Modeling Earth Systems*, *15*, e2022MS003426. https://doi.org/10.1029/2022MS003426

Received 27 SEP 2022
Accepted 17 OCT 2023

Systematic Bias Correction in Ocean Mesoscale Forecasting Using Machine Learning

Guangpeng Liu[1], Annalisa Bracco[2], and Julien Brajard[3]

[1]Department of Oceanography, School of Ocean and Earth Science and Technology, University of Hawai'i at Manoa, Honolulu, HI, USA, [2]School of Earth and Atmospheric Sciences and Ocean Science and Engineering Program, Georgia Institute of Technology, Atlanta, GA, USA, [3]Nansen Environmental and Remote Sensing Center (NERSC), Bergen, Norway

Abstract The ocean circulation is modulated by meandering currents and eddies. Forecasting their evolution is a key target of operational models, but their forecast skill remains limited. We propose a machine learning approach that improves the output of an ocean circulation model by learning and predicting its systematic biases. This method can be applied a priori to any region, and is tested in the Gulf of Mexico, where the Loop Current (LC) and the large anticyclonic eddies that detach from it are major forecasting targets. The LC dynamics are recurrent and lie on a low-dimensional dynamical attractor. Building upon the information gained analyzing this low dimensional attractor, we improve the representation of sea surface anomalies in model outputs through information from satellite altimeter data using a Sequence-to-Sequence model, which is a special class of Recurrent Neural Network. Building upon the HYCOM-NCODA analysis system, we deliver a correction to the forecast at the observation resolution. For at least 15 days the proposed method learns to forecast the systematic bias in the HYCOM-NCODA, outperforming persistence, and improving the forecast. This data-driven approach is fast and can be implemented as an added step to any dynamical hindcasting or forecasting model. It offers an interesting avenue for further developing hybrid modeling tools. In these tools, fundamental physical conservations are preserved through the integration of partial differential equations which obey them. In addition, the method highlights specific deficiencies of the hindcast system that deserve further investigation in the future.

Plain Language Summary Predicting the evolution of ocean circulation using a physically constrained models is critical for weather forecasting and for quantifying nutrient transport and water mass exchanges, but models are far from perfect. We propose a generalized data-driven method to improve a circulation model performance by identifying and predicting its systematic biases, and we test it in the Gulf of Mexico. We use a recurrent neural network to improve and predict the evolution of sea surface anomalies over 7–15 days using information obtained from satellite altimetry data. The results show that our proposed data-driven method is fast and accurate and can be easily implemented as an additional step to other dynamic models, providing an interesting avenue for further development of hybrid forecasting tools.

1. Introduction

The application of machine learning (ML) algorithms to various aspects of weather and climate forecasting has received increasing attention over the past decade. In the weather arena, ML has been used in concert with output from numerical weather prediction models to improve forecasts (Chapman et al., 2019; Davò et al., 2016; Rasp & Lerch, 2018), predict extreme events (e.g., Chattopadhyay et al., 2020) or assess weather risk (McGovern et al., 2017). In the climate community, much work has been devoted to downscale general circulation model simulations to higher horizontal resolution (e.g., Rodrigues et al., 2018), to identify low resolution modeled states conductive to extreme events in the observations and future projections (Herman & Schumacher, 2018; Kurth et al., 2019; Lagerquist et al., 2019; Larraondo et al., 2019), and to improve parameterizations of physical processes (Brenowitz & Bretherton, 2018; Rasp et al., 2018; Zanna & Bolton, 2020). Purely ML-based, data-based models have been built as well, and used to address open questions in weather and climate science (e.g., Dueben & Bauer, 2018; Scher & Messori, 2019; Weyn et al., 2019). Attention has been paid also to forecasting the ocean evolution, which is slower than that of the atmosphere, but may have societal relevance as much as weather. One example is the work of Wang et al. (2019), who developed a Recurrent Neural Network (RNN) to predict the ocean circulation in the Gulf of Mexico. The network is trained on model data from the HYCOM Navy Coupled Ocean Data Assimilation (NCODA) analysis (Cummings & Smedstad, 2013) and predicts the

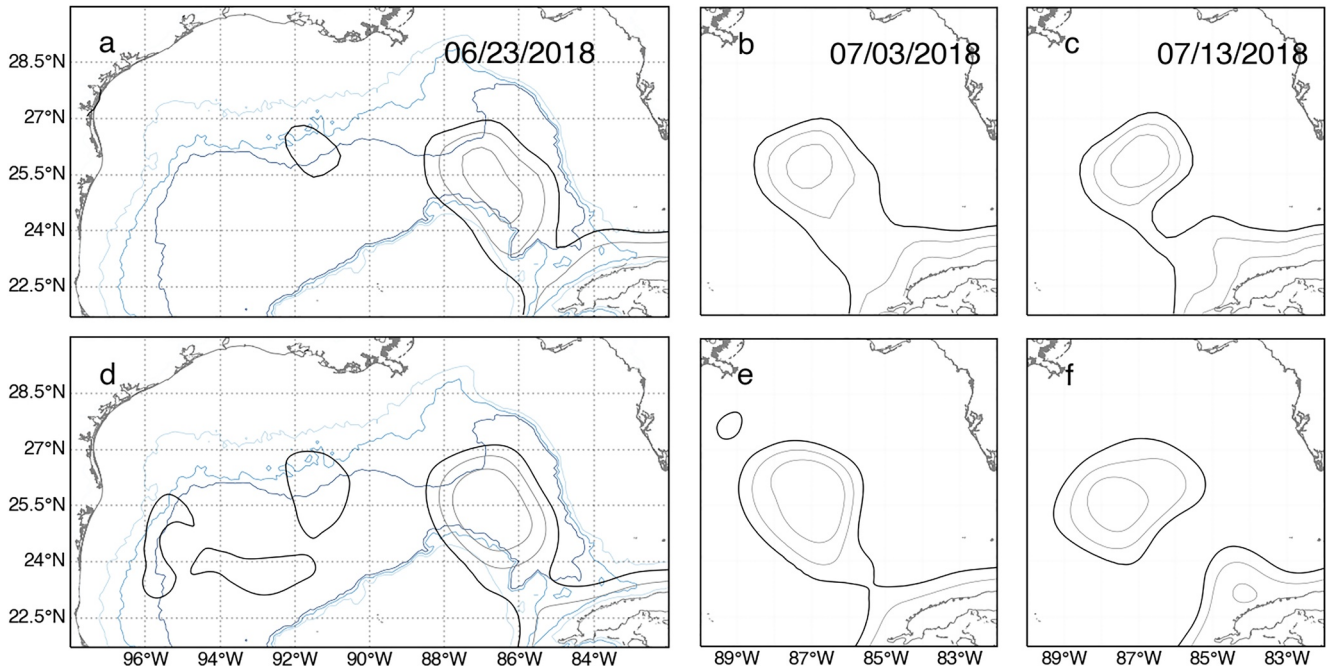

Figure 1. Domain (a and d) and sea surface height anomaly (SSHa) showing a detachment event in 2018 (GCOOS, top panels; HYCOM, bottom panels). Blue curves show 1,000 m, 2,000 m, and 3,000 m bathymetry. Black contours are 0.17 m (thick), 0.4 and 0.6 m SSHa.

evolution of the upper circulation in the Gulf over several weeks, with greater skill than the analysis. However, a common, serious problem when predicting temporal evolution using ML methods alone, is that physical relationships and constraints may be violated. Additionally, the RNN in Wang et al. (2019) is based on a modeling system which has a systematic bias, as noted by Liu et al. (2021). Here we propose a ML approach that circumvents both issues and builds upon dynamical system theory. We show that we can improve the ocean mesoscale hindcasting/forecasting obtained through an ocean circulation model at a minimal computational cost using satellite data as added information to the ML system. We apply it, as Wang et al. (2019), to the Gulf of Mexico, where many people work in the oil and fishery industry and live in low-lying coastal areas often traversed by tropical storms and hurricanes, and ocean hindcasting and forecasting are very important.

The circulation of the Gulf of Mexico is dominated by large mesoscale structures. The Loop Current (LC), which enters the Gulf from the Yucatan Channel and exits it through the Florida Straits, is approximately 250 km wide and 1,200 m deep, and bring warm and salty Atlantic water into the basin, and the Loop Current eddies (LCEs), shed from the LC at irregular intervals have size of 200 km in diameter (Hamilton, 1990). The interval at which the LCEs are shed varies between half a month to more than a year (Hall & Leben, 2016). After their formation, they migrate slowly westward and are dissipated through interactions with the west shelves. The domain, current system, and an example of LCE formation are presented in Figure 1. Forecasting the LC and LCEs evolution is fundamental to predicting tropical storms and hurricane intensity given that the LC system carries large heat anomalies (e.g., Jaimes et al., 2016), and impacts the transport of nutrients, freshwater and pollutants in the basin (e.g., Liu et al., 2018, 2022) as well, with relevance to ecosystem services, given the relatively low nutrient content of the water carried by the LC system.

In Liu et al. (2021) we adopted a dynamical system approach to describe the mesoscale motions in the Gulf of Mexico in the sea surface height (SSH) field and explored the local dimension of the low-dimensional attractor (Lorenz, 1980), following previous works (Falasca & Bracco, 2022; Faranda et al., 2017; Lucarini et al., 2016). We compared this metric in the GCOOS (Gulf of Mexico Coastal Ocean Observing System) altimeter derived SSHa data and in the HYCOM analysis over 16 years, finding that most of the time less than 10 degrees of freedom are sufficient to describe the average dimension of the LC attractor. We also showed that the LC phase-space trajectory was on average more stable in the observations (GCOOS) than in the model (HYCOM) product even when the HYCOM output was downscaled to GCOOS resolution. This is shown in Figure 2, where the two-dimensional projection of the (higher dimensional) attractor is plotted over the period considered in this

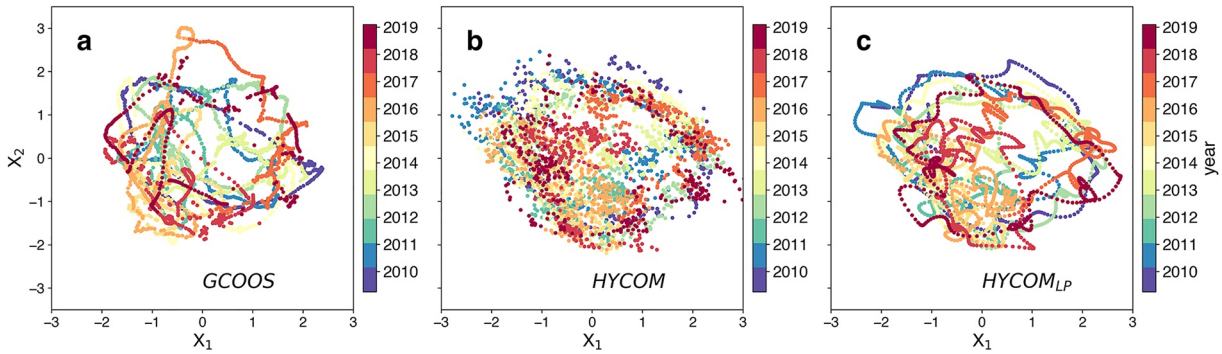

Figure 2. Time evolution of the LC attractor (calculated from SSHa) shown as two-dimensional phase-space projection of the first two principal components in the GCOOS (a), HYCOM (b) and 14-day low-pass filtered HYCOM (c).

work. The discrepancy among the attractors, which could be due in part to the temporal processing and spatial interpolation of the satellite data, was quantified in terms of local dimension in Liu et al. (2021) and points also to the common knowledge that forecast/hindcast models and ocean circulation models in general are systematically biased and tend to overestimate the rate at which LCEs form in the Gulf of Mexico (NASEM, 2018). In other words, the LC is more stable to eddy detachments than shown by HYCOM and more generally by ocean-only models. Following the evaluation of the local dimension of the attractor, we suggested that a dynamical system approach coupled to ML could inform operational ocean forecasts about the bias and help improving forecasting skill. In this work, we delve into how such improvement may be achieved.

Best results are obtained using a Sequence to Sequence (seq2seq) model (Cho, van Merriënboer, Gulcehre, et al., 2014; Sutskever et al., 2014), which is a special class of Recurrent Neural Network (RNN) architectures, consisting of two Gated Recurring Units (GRU) cells. As in Wang et al. (2019), we focus on SSHa, a variable of great importance in operational ocean forecasting, and use again the HYCOM-NCODA analysis (indicated as HYCOM in short in the following) and an altimetry-based data set in lieu of observations. The RNN model is applied to HYCOM and GCOOS to obtain corrected forecast at the (lower) resolution of the altimetry data set, and could be followed by a conventional spatial interpolation method or a super-resolution convolutional neural network (SRCNN) for resolution conversion back to the analysis grid. In simple terms, we use ML to characterize and learn the model bias, that can be then used to correct the model output improving predictability skill. The method we propose is general and efficient, and while applied here to only one variable, it could be easily extended to include multiple fields, yielding even better outcomes, or extended to atmospheric and climate models.

The proposed method is potentially generalizable to any situation where a systematic bias is identified through a comparison between model outputs to reanalysis or (gridded) observations. Its relevance is therefore broader than the specific case addressed here, and independent of using GCOOS data as proxy for the true ocean.

The remainder of this paper is organized as follows: in Section 2, we review a recent analysis of the attractor describing the SSHa evolution in the Gulf of Mexico in GCOOS and HYCOM, the main characteristics of the data sets used, and their major differences in terms of attractor. We also detail the new ML based SSHa forecasting method. Results and evaluation are presented in Section 3. Conclusions and discussion follow in Section 4.

2. Method

2.1. SSH Data

We focus on SSH anomalies, which are representative of the large-scale mesoscale circulation. In GCOOS four altimetry satellites are merged into daily gridded maps following Leben et al. (2010). GCOOS SSHa data have a spatial resolution of $0.25° \times 0.25°$ and are obtained by applying objective mapping to along-track satellite altimetry, then gridding uniformly the outcome with daily sampling. As a result, the effective spatiotemporal resolution of the gridded product is lower than its nominal one. Additionally, the orbital characteristics and repeat cycles of altimeters, which in the Gulf of Mexico vary between 10 and 35 days, suppress in part temporal variations

up to about 10 to 14 days (see Ballarotta et al., 2019 for a general discussion of this problem). The model data set is the HYCOM-NCODA analysis available with 3-hourly frequency on a $0.04° \times 0.04°$ horizontal resolution grid. The analysis is produced by the Naval Oceanographic Office together with a real-time, widely used forecast for the Gulf of Mexico. The analysis system assimilates along-track satellite altimeter obtained from the Naval Oceanographic Office Altimeter Data Fusion Center, SST satellite observations and all available in situ temperature and salinity profiles using the NCODA system. Following repeated upgrades in the forecasting system, only limited periods of the analysis have been run for any given version of HYCOM-NCODA (HYCOM in short). The period analyzed here is 1 April 2009, to 2 July 2019 and we consider the experiments GOMl0.04/expt_31.0 and GOMl0.04/expt_32.5, available respectively from April 2009 to July 2014 and from April 2014 to February 2019 which have the same horizontal resolution (we use expt_32.5 for the overlapping months). The analysis is run in real time, assimilates SST and profiles using FGAT (first guess at appropriate time) and is forced by atmospheric fields from the Center for Ocean-Atmospheric Prediction Studies (COAPS). HYCOM expt_31.0 (expt_32.5) performs four (one) days of analysis using all observations received since the previous analysis and generates a forecast for the subsequent 7 days. A major difference between these two experiments is the number of vertical layers, 20 in expt_31.0 and 27 in expt_32.5, respectively. We verified that SSHa differences between the two experiments are small during the overlap time period. Note that although HYCOM has rapid version development, the latest versions, such as GOMl0.04, are not yet available as analysis product dating back at least a decade for the required training. Therefore, we cannot assess the performance of analysis performed with newer configurations or more vertical layers.

We stress that we assume the altimetry-based GCOOS data set to be representative of the true ocean state, despite the limitations of this product, linked especially to the effective time and space resolution of the data used to generate it. Despite the relatively low spatial and temporal resolution of GCOOS, the geostrophic currents derived from the satellite product have been often found more accurate than those derived from data assimilative models, for example, when used to hindcast satellite-tracked drifter trajectories in the eastern Gulf of Mexico during spring and summer 2010 following the Deepwater Horizon blowout (Liu et al., 2014). Altimetry-derived products (SSHa and geostrophic currents) have been shown to be as, and at times more accurate than the HYCOM analysis and in general data-assimilative models also in relation to the LC patterns (Alvera-Azcárate et al., 2009; Y. Liu et al., 2013, 2016; Weisberg & Liu, 2017). In addition, the method we propose is general and independent on the specific choice of GCOOS as reference truth.

2.2. PCA

Our overarching goal is to develop a method that learns the model bias by comparing true and modeled fields (while recognizing GCOOS limitations), and then bias-corrects the model output. We want to do so using as little computational time as possible, to eventually introduce the bias-correction step in a real-time prediction system such as HYCOM-NCODA. To this end we adopt the dimensionality reduction technique, principal component analysis (PCA) already used in Liu et al. (2021) to characterize the LC attractor. PCA is one of the most popular multivariate statistical techniques in climate science. Also known as empirical orthogonal function (EOF) analysis, PCA is a linear dimensionality reduction technique that allows to transform data from a high-dimensional space into a lower dimensional space which still contains most of the original information (e.g., Abdi & Williams, 2010). In brief, PCA is a decomposition of an $m \times n$ matrix $\mathbf{X} = \mathbf{F}\mathbf{Q}^T$. $\mathbf{Q}$ is the loading matrix ($n \times n$), an orthonormal projection matrix such that $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$ and $\mathbf{F}$, the score matrix ($m \times n$), is the projection of $\mathbf{X}$ onto the new space. $\mathbf{Q}$ is calculated from the covariance matrix $\mathbf{C}$ given $\mathbf{Q}\Lambda\mathbf{Q}^T = \mathbf{C}$, where Λ is the diagonal matrix of eigenvalues. Finally, the first k columns of $\mathbf{Q}$ are selected according to their relative importance derived from $trace(\mathbf{C})$ and are used to reconstruct $\mathbf{X}_k$ spanned by the first k basis vectors that contain most of the data variance through linear combination. Hereafter the decomposed spatial (loading matrix $\mathbf{Q}$) and temporal components ($\mathbf{F}$) of the SSHa fields are referred to as EOFs and PCs. PCA is performed on HYCOM and GCOOS separately to obtain the dominant leading modes of each in the GoM. In this work, we perform all the correction and prediction processes on PC time series, and re-build the SSHa field using EOFs as last step.

A practical question that arises is how many modes should be retained. Here, we chose to retain the first 10 modes that explain 78.25% of total variance in HYCOM and 72.93% in GCOOS, respectively, based on the quantification of the attractor's dimension in Liu et al. (2021). We acknowledge that with this choice we neglect higher modes which may contain processes that are potentially crucial for nonlinear interactions within the system. In

a practical application of our method, however, these higher modes, while not bias-corrected, would continue to be included in the modeled forecast. Alternatively, the adoption of kernel functions, instead of PCs, would capture additional nonlinearities, but with the major disadvantage of proving only rough estimations during back-projection, making the reconstruction process more difficult.

To quantify the role of the missing information given our choice of using the first 10 PCs, we compared outcomes using our proposed method versus training on the complete raw data set. We found that performances were comparable in their root mean square errors (RMSEs) on the 15th day prediction (see further below for details), with values of 0.083 and 0.085 for PCA reconstructed (first 10 PCs) and raw-data based methods, respectively. The advantage of the PCA-based method is that it is noticeably faster to train.

In our analysis of the Loop Current attractor in Liu et al. (2021), we pointed out that significant differences between HYCOM and GCOOS emerge in the second mode. Here we examine again the two data sets over the whole Gulf for a different -and shorter—time period, confirming this result (Figure 2): the patterns of the second modes do not correspond. Previous studies have shown that there is a relationship between the propagating characteristics of the LC eddies and the leading modes of SSHa variability in the GoM. In particular, Chang and Oey (2013), using AVISO data, associated the growth and retraction of the LC to the first mode, and its west-northwestward expansion and subsequent eddy shedding to the second. Patterns in Chang and Oey (2013) are consistent with those of GCOOS, despite the different time considered, but not with HYCOM, where the second mode appears to describe the LC expansion to the northwest. The HYCOM behavior is verified also if a 2D low-pass filter (40 km) is applied to the HYCOM data set. The low-pass filtered HYCOM modes are indistinguishable from those in Figure 3, and highly correlated (correlation coefficients are always greater than 0.99).

Given the discrepancy, we select for each GCOOS mode, the HYCOM one with the highest PC correlation coefficient, and the corresponding spatial EOFs are checked visually. The HYCOM modes are then reordered as shown in Table 1 to fit the GCOOS modes. The temporal PC correlations of all the reordered modes are high (ranging from 0.48 for GCOOS PC7 to 0.83 for PC4) except for the second PC in HYCOM, which does not have a matching mode in GCOOS and is shown in the table with its correlation value with the 10th GCOOS PC. We exclude this mode (and the 10th GCOOS mode) from the following calculations and use only the 9 remaining leading modes. The second HYCOM mode, however, contains a crucial part of the spatiotemporal variability in the model, explaining 17.72% of the variance. Finding its related observational information in GCOOS may help understanding the predictive skill of HYCOM and its systematic bias. HYCOM PC2 is correlated to the first three PCs of GCOOS with coefficients 0.59, 0.44, 0.42, respectively, while the highest coefficient with any of the remaining PCs is 0.12. In addition, we found that the spatial pattern of the HYCOM 2nd mode is very similar to the pattern of the standard deviation (STD) of the difference between of the sum of the first three GCOOS modes and the sum of the first, third and fourth modes in HYCOM. The respective PC time series, however, are not correlated when considered simultaneously (c.c. = −0.01), but their correlation coefficient is as large as 0.73 if a lag of 422 days is considered. We have no definitive explanation for the model behavior, which deserves further evaluation beyond the scope of this work. We hypothesize that this result is due to either (smoother) GCOOS not being representative of the actual truth, or, most likely, to the second mode of HYCOM being incorrectly derived from the first three modes of GCOOS due to some bias introduced by the assimilation algorithm. In any case, we exclude the second mode of HYCOM, which results in information being removed in the ML calculations. From the ML perspective, doing so is comparable to introducing noise in the training data and might impact the method's ability to predict the bias accurately. The impact will, however, be reflected in the ML model error.

2.3. Data Processing

Unlike a traditional correction or prediction task with clean data sources, the characteristics of the GCOOS data should be accounted for. GCOOS necessarily smooths the satellite altimeter data to account for orbital characteristics and satellite repeat cycles. Therefore, as first step we resample the GCOOS and HYCOM SSHa to daily average, remove from each data set its global mean, both spatial and temporal, and low-pass filter both data sets using a Butterworth filter with a 14-day cutoff frequency (see Figure 2 for a comparison of the phase space attractor in the two data sets once the filter is applied). Second, we perform the PCA analysis to get the corresponding PCs, and use the z-score normalization, defined as $X = (x - \mu)/\sigma$, where x is the original data set, μ and σ are its mean and standard deviation, respectively, to standardize the PC time series. This produces a new data set with a mean of zero and a standard deviation of one. These two steps were undertaken to generate Figure 3,

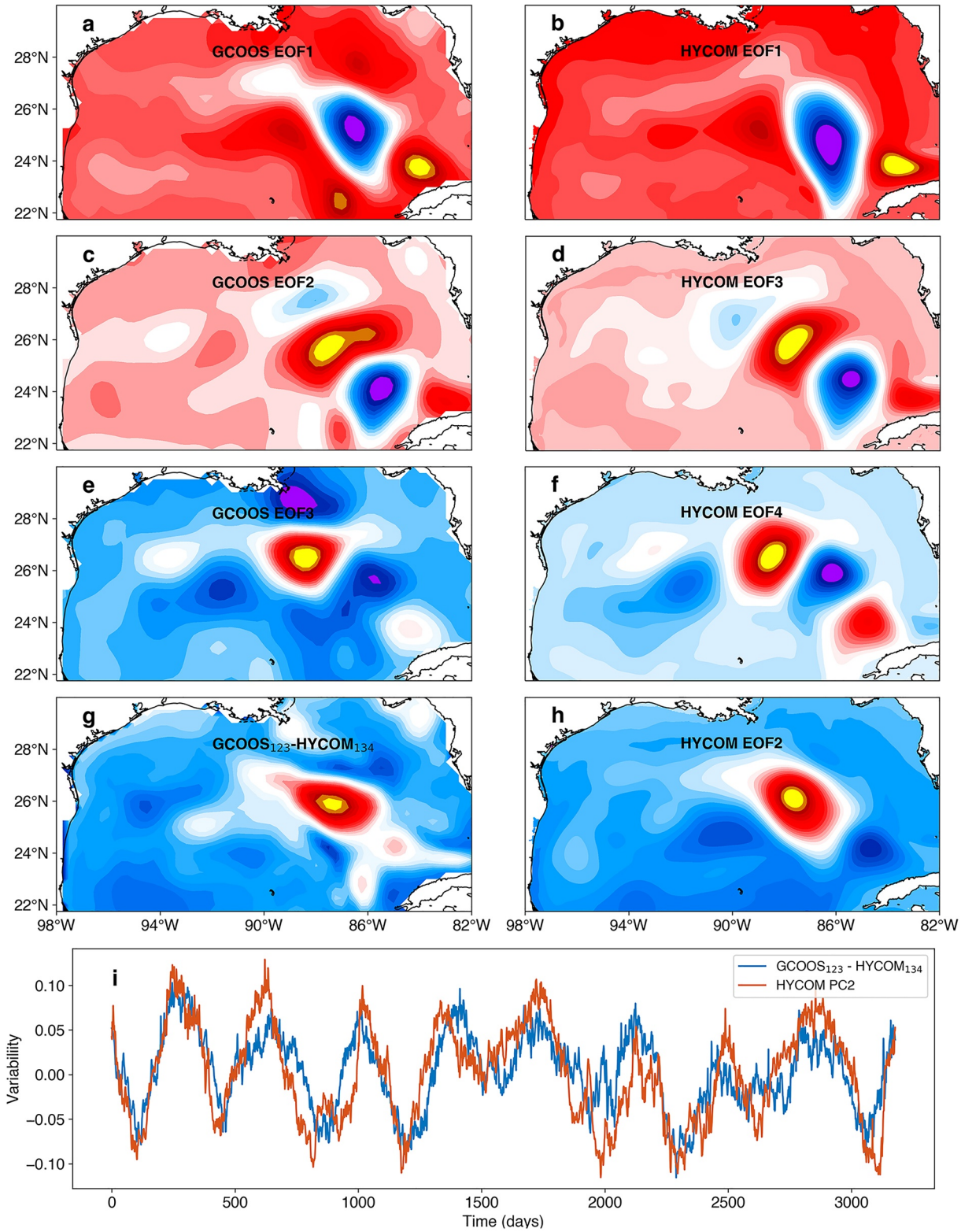

Figure 3. First three EOF modes for GCOOS (a, c, and e), and first, third and fourth in HYCOM (b, d, and f); g is the standard deviation of the sum of the first three modes of GCOOS minus the sum of modes (1 + 3 + 4) in HYCOM, while h is the second EOF of HYCOM, and their time series (domain averaged) are shown underneath (i) with a lag of 422 days. See Section 2.3. For a description of the data processing.

Table 1
Re-Ordered GCOOS and HYCOM Modes, Their Correlation Coefficient and Percentage of Variances

GCOOS Mode	1	2	3	4	5	6	7	8	9	10
	19.1%	15.8%	12.0%	6.3%	5.2%	3.7%	3.2%	2.7%	2.6%	2.3%
HYCOM Mode	1	2	3	4	5	6	7	8	9	10
	24.5%	3 9.8%	4 8.6%	5 4.8%	6 3.7%	7 2.7%	10 2.0%	8 2.3%	9 2.1%	2 17.2%
Correlation	0.76	0.68	0.82	0.83	0.73	0.70	0.48	0.66	0.60	0.11

but indistinguishable plots and nearly identical correlations can be found using the original data resampled to daily averages. The filtering manipulates the portion of the model data that should have comparable evolution in phase-space, a process commonly known in machine learning as feature engineering. Third, we use a min-max normalization to set the data sets to vary between −1 and 1, following $X = 2 * (x - min(x))/(max(x)\text{-}min(x)) - 1$. Usually, using either the previous z-score or a min-max normalization is sufficient for model training. Here we chose to combine them because the results so obtained were slightly better. Next, we split the resulting time series into three groups, training, validation and testing, with the first group, ∼80% of the total data, from 2009-05-21 to 2017-02-18 (2,830 samples) being used to update and optimize the model parameters, the validation set of about ∼10% of the data from 2017-02-19 to 2018-02-13 (360 samples) being used to select the best performing model, and the last testing group including 340 samples (2018-02-14 to 2019-01-19) used to evaluate the performance of the proposed approach.

The final step consists in preparing input-output pairs for the following model training and prediction. Let us denote x_k^a the HYCOM PCs and x_k^o GCOOS PCs at time t_k. The goal of our model is to input the previous m days of HYCOM $x_{k-m:k}^a$ and GCOOS $x_{k-m:k}^o$ data and output the analysis corrected by the satellite information for the next n days. To do so, we use the difference between them as the target $y_{k+1:k+n} = x_{k+1:k+n}^a - x_{k+1:k+n}^o$ and all the available known information from the past m days as input features. As illustrated in Figure 4, for each data set we use a sliding window with a stride of one to select sequential samples, with an input sequence having previous known information (blue shading) and its corresponding output sequence (green shading). For example, using 14 and 7 days for m and n, respectively, the dimension of the input training data is 2,830 samples (about 80% of the total data, each sample is an input-output sequence pair), 14 previous days and 18 time series (9 GCOOS PCs and 9 corresponding HYCOM PCs, namely input features), and the output size is 2,830 samples, 7 predicted days and 9 time series (9 bias time series). For the testing data, we generate two sets with two different low-pass filtering strategies. We define the first strategy as *realistic* because, for a given time t_k, the low-pass filtering is applied only to data known from the past (i.e., the input sequence) as future observations (i.e., the output sequence) are not available. In this case the prediction error includes model error and data error due to the filtering process. In the second strategy, we low-pass filter both input and output sequences given that all information is available to us. This *ideal* strategy would apply to other situations in which no filtering is required and input and output data share the same time resolution and the prediction error would contain only model error.

2.4. Prediction Model

For the multivariate (9 PCs) multi-task (n predicted days) time series prediction problem at hand, we choose to use the sequence to sequence (seq2seq) model, also known as the Encoder-Decoder model, and compare the results with a simpler ridge regression model, a linear regression model that uses several past time steps as input features. The seq2seq model (Figure 4) was first introduced for machine translation (Cho, van Merriënboer, Gulcehre, et al., 2014; Sutskever et al., 2014) and has been since widely applied in fields such as speech recognition, image captioning, conversational models, or text summarization. In seq2seq learning, a recurrent neural network (RNN) based model is trained to map an input sequence (words, letters, time series, etc.) to an output sequence of items. The input and output do not have to be of the same length. The seq2seq model contains two RNN structures, for example, Long Short-Term Memory (LSTMs), or Gated Recurrent Units (GRUs) that are treated as the encoder and decoder cell. We chose the GRU model since it has a simpler model structure and, in general, has a similar predictive capability to LSTM (we tested the LSTM as well). The encoder part converts the given input sequence to a fixed length vector $\boldsymbol{C}$, which captures the context of the input sequence in the form of a hidden state vector.

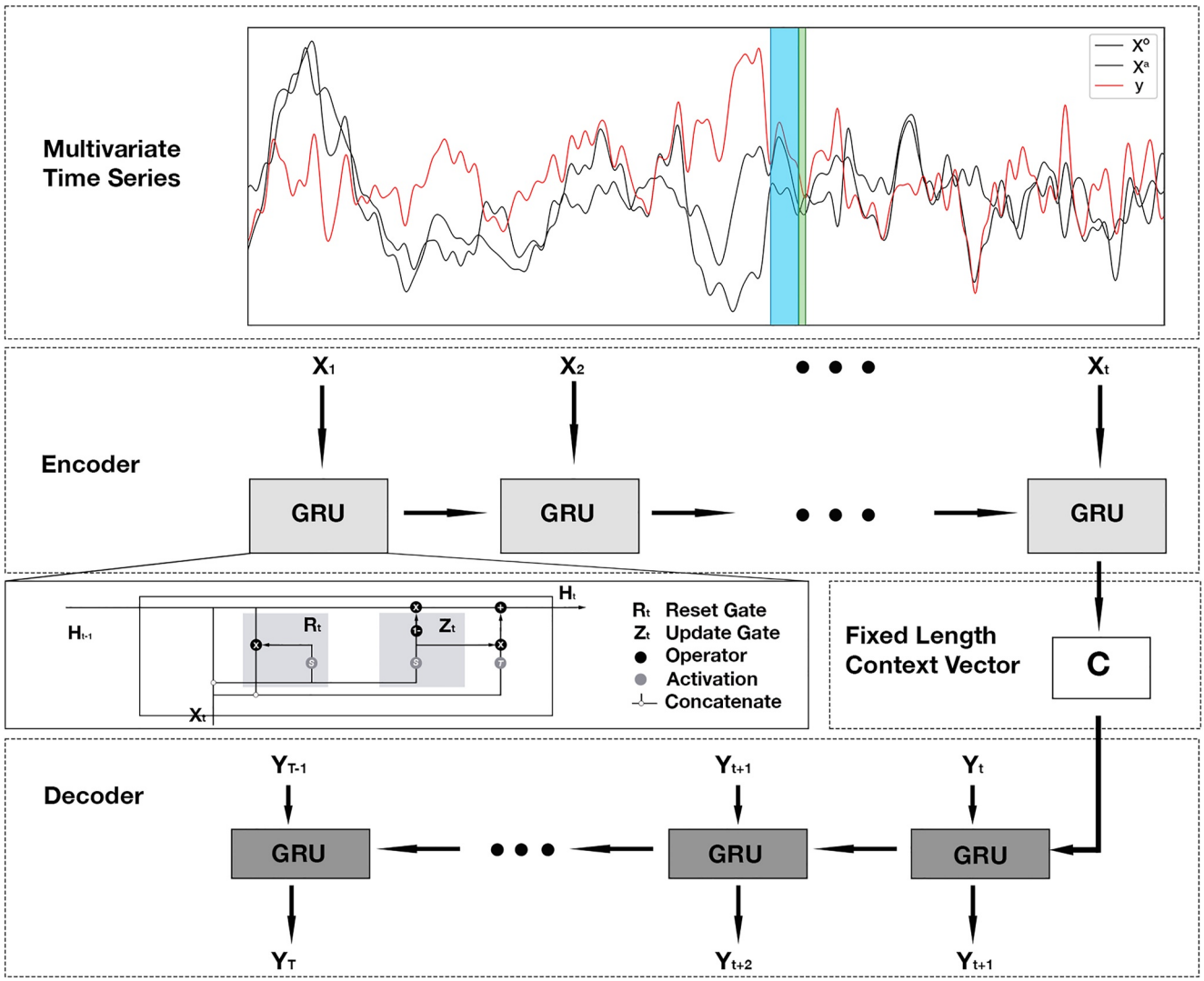

Figure 4. Schematic diagram showing the prediction pipeline. For sequences with m inputs and n outputs at time k, we define the input sequence in the Encoder unit as $\mathbf{X}$ $[x^a_{k-m:k}; y_{k-m:k}]$, and the Decoder prediction as $y_{k+1:k+n}$. GRU and $\boldsymbol{C}$ are Gated Recurrent Units and context vector, respectively.

The context vector $\boldsymbol{C}$ is then sent as input to the decoder to produce the output sequence, and after every successive prediction, it uses the previous hidden state to predict the next instance of the sequence. The GRU encoder and decoder have different weights. The GRUs by design take two inputs: the current state and a representation of the previous state. Therefore, the output at time step t_k depends on the current input as well as the input at time t_{k-1}. The sequential information is preserved in a hidden state of the network and used in the next instance, and this is the reason they perform well when posed with sequence related tasks. More details of GRUs and the coupled gating mechanism can be found in Cho, van Merriënboer, Bahdanau, and Bengio (2014).

In this study, the past m-day information of HYCOM PCs $x^a_{k-m:k}$ and GCOOS PCs $x^o_{k-m:k}$ is stored in the context vector $\boldsymbol{C}$ after being processed through the encoder and is passed to the decoder to predict the future n-day bias, $y_{k+1:k+n}$, recursively. Finally, the n predicted biases are applied to the n days of HYCOM PCs to obtain the corrected forecast for the subsequent SSHa reconstruction.

2.5. Training

Summarizing, we first shuffle the training samples to break down the long-term sequential information, then we feed the model with the past 14 days HYCOM and GCOOS PCs, and predict the difference between them

Table 2
Detailed Parameters Used in the Seq2seq Neural Network and Training Process

Data size	3,530 days
Input sequence length	14 days
Output sequence length	7/11/15 days
Stride	1 day
Number of layers	1
Activation function	Tanh
Hidden cells	64
Loss function	Mean absolute error
Learning rate	$1e^{-5}$
Optimizer	Adam
Dropout rate	0.1
Training epochs	40,000
Batch size	32

Note. A stride of 1 day was used to select the moving input-output sequences for training, validation and testing data sets. Compared to other activation functions, Tanh has been tested and found to be the most suitable one for the present task. Dropout refers to shutting off some neurons with a certain probability during training. We set the dropout rate to 0.1 to add a level of uncertainty to the model.

for the next several days. The goal of the neural network is to minimize the mean absolute error (MAE) loss, which equals to $\frac{1}{N}\sum_{n=1}^{N}|y_n - x_n|$ and quantifies the difference between the truth (GCOOS) and the forecasts (HYCOM). By doing so, we correct the model forecast. For each GRU cell, we use 1 hidden layer, and 64 hidden neurons for each layer. The optimizer chosen for model training is the Adam optimization algorithm (Kingma & Ba, 2014) with a learning rate of $1e^{-5}$. The model is trained for 40,000 epochs, where the number of epochs defines the number of times that the learning algorithm goes through the entire training data set. Despite the large number of epochs, which is due to the low learning rate, the training is fast given the size of the data sets and the model used. Further details can be found in Table 2.

For the ridge regression model, we use the training set and k-fold cross-validation to obtain the optimal model, and then the testing data for predictions to compare the outcome with the seq2seq model. The k-fold cross-validation is a popular method in applied ML to estimate a model's performance on unseen data, especially when the available data is limited. In brief, we split training data into k groups (here, k equals 5), train the model on k-1 groups, and make evaluations using the remaining group. We repeat this step five times and take the average to get the final result.

3. Results

To evaluate the performance of our method, first we show the predictive ability of the seq2seq RNN model by investigating the PCs time series. Then, we assess the final product obtained by combining the predicted PCs and GCOOS EOFs.

3.1. Seq2seq Training and Prediction Performance

Figure 5 shows the loss evolution of the seq2seq model for the 7-day forecast case for the training and validation data sets. The higher validation loss (0.014 higher at epoch 38,246) reflects a typical overfitting issue and is likely due to the partitioning regime that makes all the validation set unseen to the model. Another possible reason is that our training data may have higher quality than validation, which makes the learning process easier, or the two may have different characteristics. When the training data set is determined, it is customary to choose simpler models or carefully tune the model parameters to tackle overfitting issues. The results presented here are after fine-tuning the model parameters. Both training and validation losses are gradually decreasing, and convergence

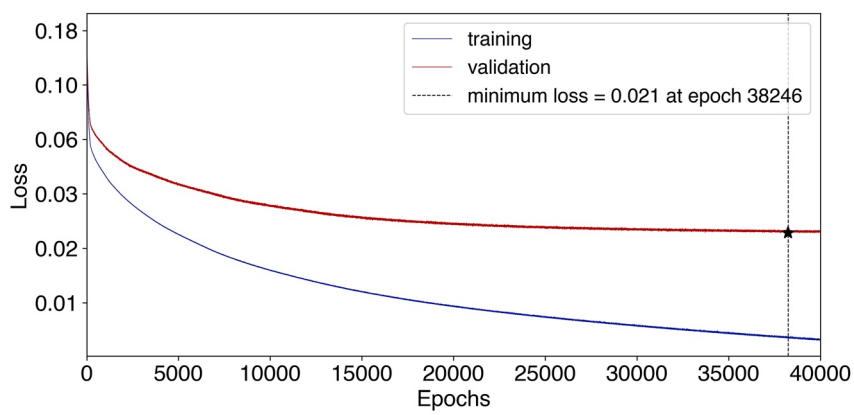

Figure 5. Training and validation loss of the 7-day seq2seq model in log scale. The black star marker and dash vertical line indicates where the optimal parameters are obtained. The tick marks for the y-axis (Loss) are converted to linear scale for immediate readability.

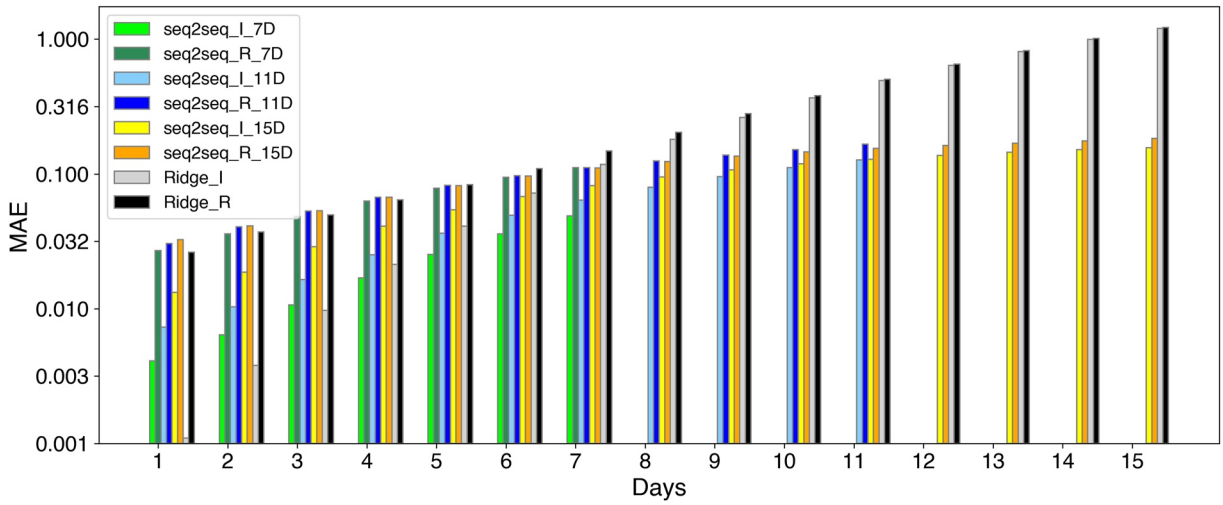

Figure 6. Bar plot of the MAE error calculated between the true and predicted PCs of 7-day, 11-day, and 15-day predictions for the base Ridge and the seq2seq models (in log scale). "I," "R," and "D" in the legend indicate "ideal," "realistic," and "prediction horizon in days," respectively.

is achieved after 38,246 epochs with a validation loss value of about 0.02. The neural network status at this iteration is saved as the optimal model parameters.

We then run the seq2seq model in prediction mode for both ideal and realistic strategies. Here, we used testing data from nearly a year (340 days) spanning from 2018-02-14 to 2019-01-19 for predictions. These data sets are unseen to the model, that is, they have not been used for either training model parameters (training data set) or selecting the best performing model (validation data set).The optimal scenario is simulated in the ideal case, where the temporal frequencies of the model and the observation are the same, while the realistic case involves applying low-pass filtering to the input PC sequences. Results are summarized in Figure 6 for 7-, 11-, and 15-day prediction. For the 7-day prediction case, the mean absolute prediction error or MAE in the ideal case (0.0212) is close to the validation loss of 0.0209 and is lower than the mean loss of the Ridge base model (0.0381), as to be expected. The MAE was also our metric of choice for model training. Not surprising, the prediction error on day 7 is significantly higher than that on day 1. Overall, the seq2seq model outperforms the Ridge model after about 3 days but is slightly worse in the first few days. This is because for RNN models, each update to the model parameters aims to minimize the total loss in the sequence data (in our case the total loss over the 7-, 11-, or 15-day predicted). The ridge regression model defined previously is used as base model. It is better suited for short-term prediction because exploits the autocorrelation of the SSHa data, but effectively predicts the first day and then recursively predicts the following days, diverging rapidly. The base model outperforms the seq2seq model on short-term (e.g., 1–3 days) prediction. This conclusion has maintained a consistency in the variability in nearly a year of testing prediction time (Figure S1 in Supporting Information S1).

Figure 7 compares the base Ridge and seq2seq 11-day predictions and truth for each PC in the ideal and realistic cases. Results with different prediction periods can be found in the Supporting Information S1 (Figures S3 and S4 in Supporting Information S1). The first three PCs have a higher standard deviation, but their predictions outperform those of the other PCs in terms of correlation and RMSD. The error for the ideal case is minimal among all cases, indicating that the low-pass filtering is the primary cause of inaccuracy and that the predictive capacity of the method we propose is high. It is worth noting that the base model works better for short-term predictions (Figure S2 in Supporting Information S1), especially for the first three modes, but performs poorly for more complex long-term predictions and for other PCs (Figure 7; Figure S3 in Supporting Information S1).

3.2. SSHa Prediction

To evaluate the improvement associated with the proposed method, in Figure 8 we quantify both the skill of the HYCOM analysis using the GCOOS data as a reference (raw HYCOM data vs. GCOOS) and the truncation error introduced by considering only the 9 modes reconstructed SSHa (reconstructed GCOOS vs. raw GCOOS). The error in panel b is determined by the number of selected modes and is thus controllable. The improved skill introduced by the seq2seq model, also in comparison to the base Ridge model, is shown in Figure 9 for the 7th day

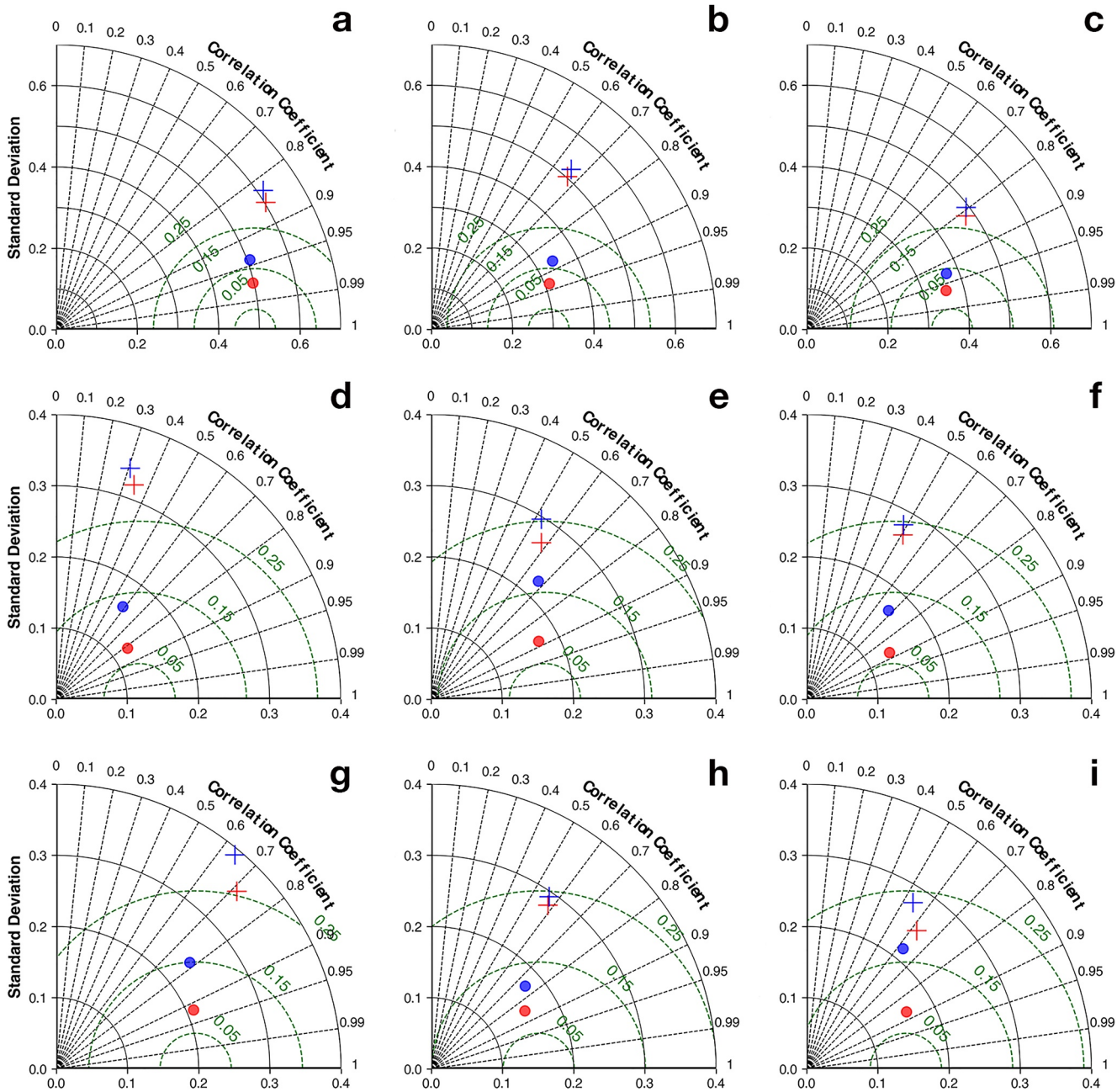

Figure 7. Taylor diagrams of the PCs (PC1 to PC9 from a to i) predictions for the 11-day case. Red and blue dots represent ideal and realistic seq2seq prediction, while the corresponding Ridge base model results are illustrated using + signs. The diagrams quantify correlation, standard deviation and RMSD (green dashed lines) between prediction and GCOOS *truth* for each PC.

prediction after converting to SSHa field (the 4th day results is shown in Figures S4a–S4d in Supporting Information S1). We compute the kernel density, which estimates the probability density function of data points based on kernels, using fields at the GCOOS resolution. That is, HYCOM fields are interpolated onto the GCOOS grid and both HYCOM and GCOOS data sets are low-pass filtered (in other words we compare the data sets used for the PCA decomposition). Both realistic and ideal cases are considered. The seq2seq model always outperforms the base model. Additionally, the model errors calculated as RMSE values are always lower than the controllable RMSE between GCOOS and the truncated GCOOS (given our choices, the true error consists of the sum of the information discarded by the PCA decomposition and the model error). This statement also holds for the 15-day prediction (Figures S4e–S4h in Supporting Information S1).

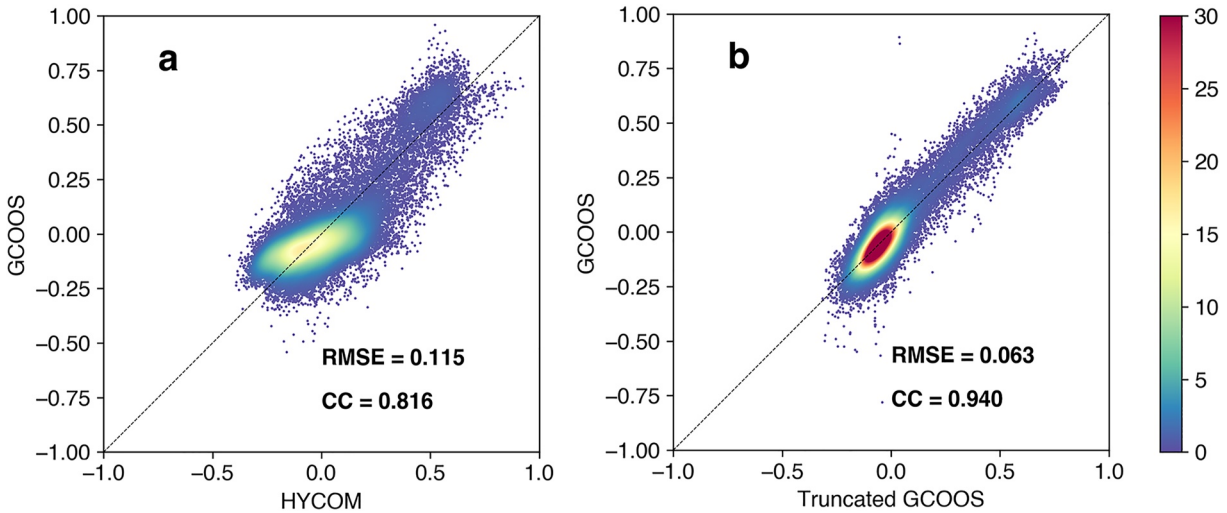

Figure 8. Scatterplots of kernel density estimation for (a) GCOOS versus HYCOM analysis, (b) original GCOOS versus truncated GCOOS (namely reconstructed using only the first 9 modes). The RMSE values are calculated using all available data, while for each subplot only 50,000 randomly sampled data points are plotted for visualization purposes.

In Figure 10 we illustrate a case of (relative) poor performance of the realistic case by selecting the 7th day prediction of the worst-case among all samples. The HYCOM analysis approximates the overall SSHa distribution but diverges from the GCOOS data in the vicinity and to the north of the LC. The GCOOS truncated SSHa (reconstructed using the first nine EOF modes) captures better the LC, but diverges in the western part of the basin where the signal is weaker than in the original (Figure 10c). The seq2seq model reproduces the target with accuracy in the ideal case (Figure 10d), but is biased in the representation of the LC shape in the realistic case (Figure 10e) due to the filtering error. Indeed, the same kind of error can be seen in the realistic Ridge model output (Figure 10f).

4. Discussion and Conclusions

In this work we present a novel approach to improve ocean hindcasting and forecasting based on bias-correcting via ML the output of a numerical model considering an observationally based data set as the truth. We apply it to the Gulf of Mexico and its SSHa field, using the HYCOM-NCODA analysis and GCOOS as model and truth, respectively. For the chosen prediction period, up to 15 days, our approach improves HYCOM skill by addressing part of the internal model error by learning and correcting for its systematic bias. This happens despite the lower spatial and temporal resolution of GCOOS that require filtering HYCOM data in the preprocessing steps. The advantage of our approach is that the outcome remains rooted in the circulation model output, which integrates a system of partial differential equations that are physically constrained, limiting the risk of obtaining physically implausible results.

The ML tool adopted is a sequence-to-sequence model, which is a special class of recurrent neural network (RNN) architectures. The RNN uses one encoder with 18 input features generated from GCOOS and HYCOM SSHa PCs. The choice of 9 PCs is dictated by a previous analysis of the SSHa attractor for the region of interest and the average number of degrees of freedom that describes it. This method could be implemented within the HYCOM-NCODA system at a low computational cost to systematically improve the forecast skill in the Gulf of Mexico, where the information is critically needed (NASEM, 2018).

Among various tests we performed on the method, we attempted to train the model with fewer data. The model error grows with less available training data, which means that the amount of input data is a strong limitation for the model performance, at least in our case. Therefore, including more fields (e.g., adding sea surface temperature information) and a longer time could potentially improve the prediction further. We also tested two others commonly used RNN techniques, namely, the attention mechanism (Niu et al., 2021) and teacher forcing (Lamb et al., 2016), and found that neither of them significantly improved the performance. This may be because only

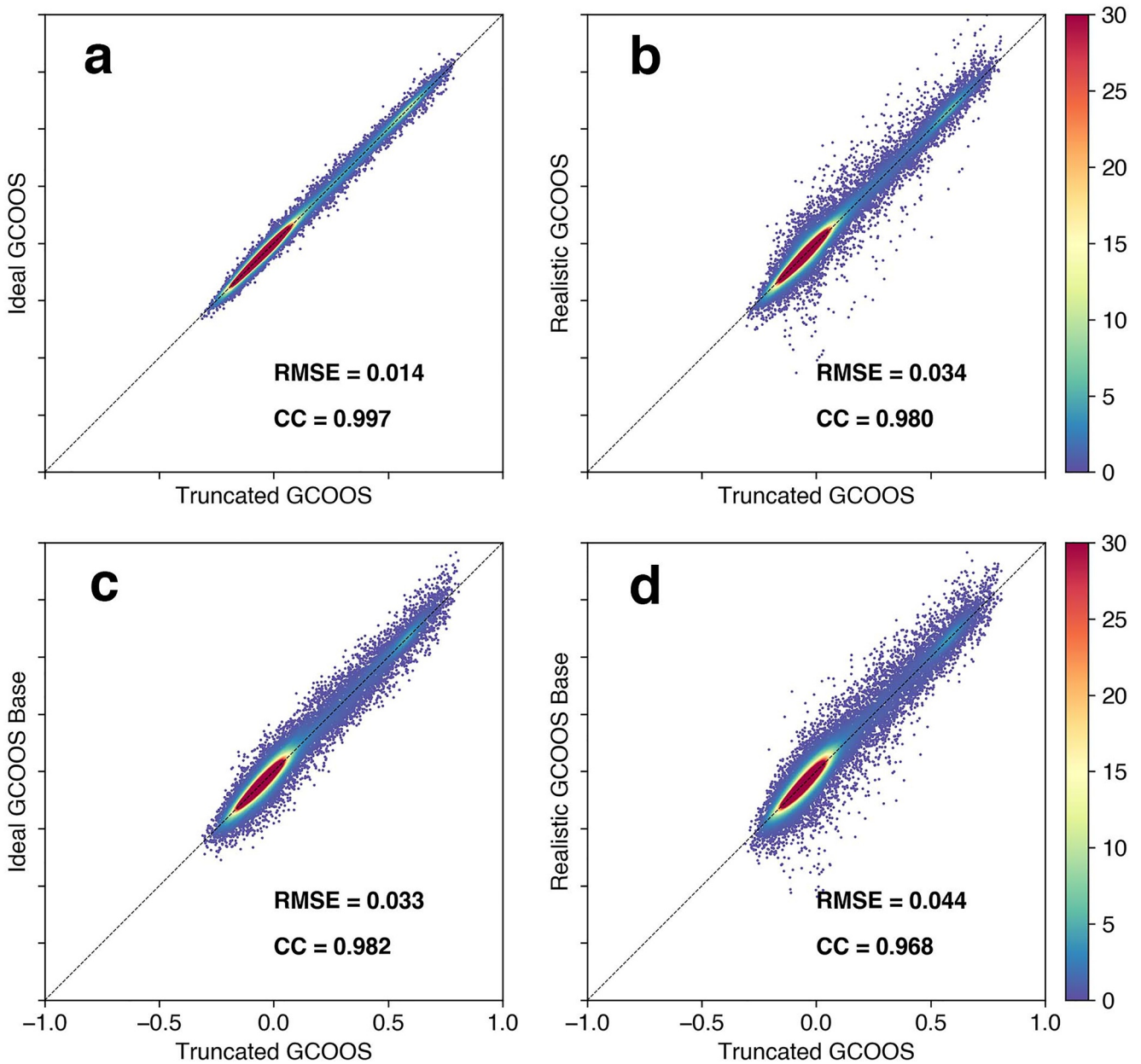

Figure 9. Scatterplots of kernel density estimation of the last day prediction for the 7-day case, (a) ideal seq2seq corrected HYCOM in GCOOS space, referred to as ideal GCOOS, (b) realistic seq2seq corrected HYCOM versus truncated GCOOS, (c) idealized ridge regression corrected HYCOM versus truncated GCOOS, and (d) realistic ridge regression corrected HYCOM versus truncated GCOOS.

one state variable and relatively short input and output sequences are used in this study, making a simple seq2seq model suitable for the job at hand.

This proposed strategy requires all PCs to have the same length; in other words, the sampling rate of observation and model must be the same (daily data in our case). In real situations observations may be sampled at a low frequency that forecasting outputs, implying that the input sequence of observations may not have the same length of the modeled one that we wish to correct. This problem can be solved by modifying the seq2seq model slightly to use two or more encoders together, that is, one encoder for daily model outputs and one encoder for the less frequently sampled observations. The outputs of these two encoders can then be processed (concatenated) by an activation layer (sigmoid function, e.g.,) or simply by linear combination to get the context vector for the decoder. This modified structure has been preliminary tested for our domain and does not affect the model performance

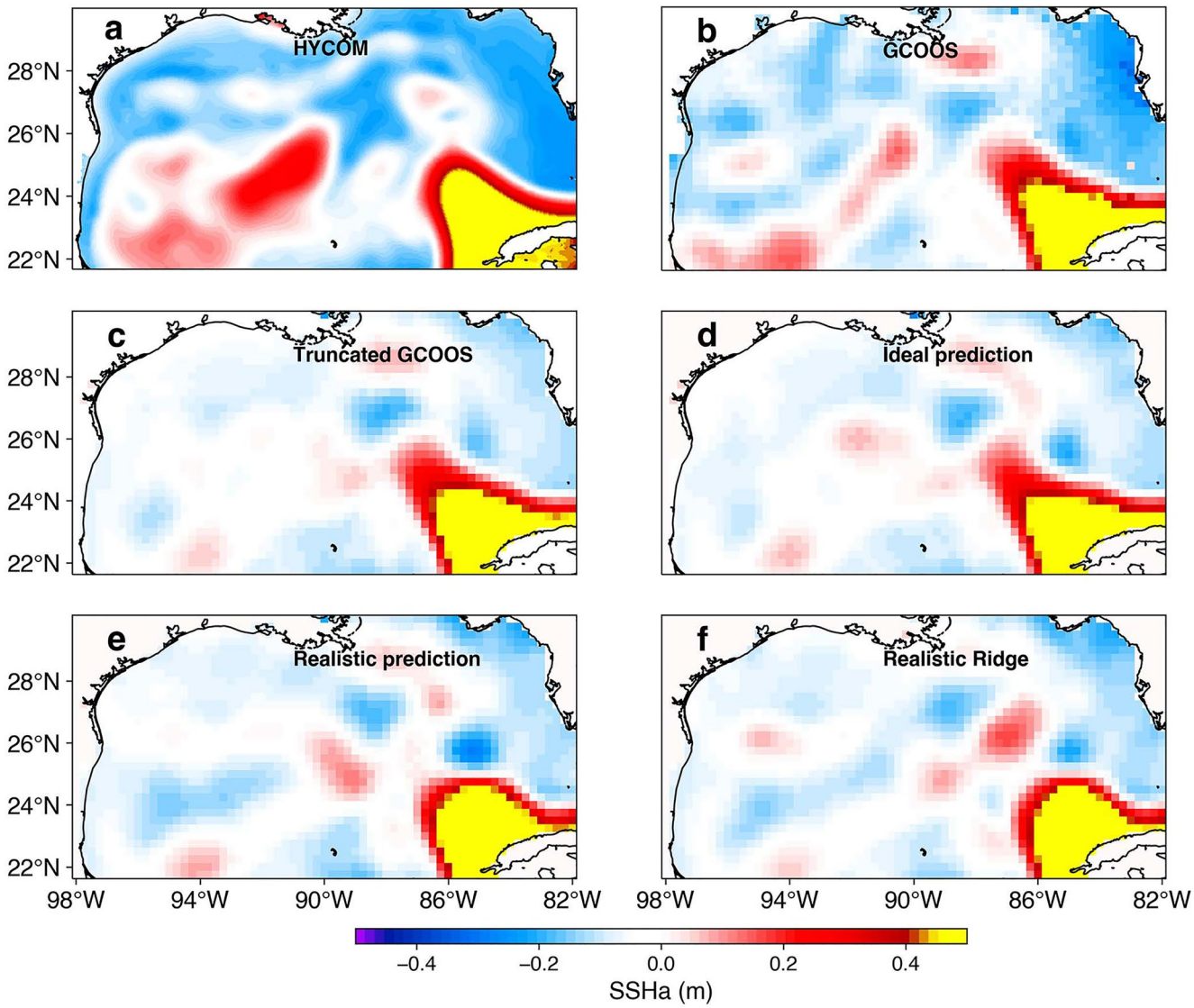

Figure 10. Worst case scenario for 7th day prediction showing the SSHa field in (a) HYCOM analysis, (b) GCOOS, (c) truncated GCOOS reconstructed from the 9 leading modes, (d) HYCOM analysis corrected with the ideal seq2seq model, (e) same as (d) but for the realistic seq2seq, and (f) HYCOM corrected using the realistic Ridge model.

in a significant way (not shown). For more complicated situations or a system with multiple fields available on different grids, for example, in situations where temperature, winds, land masses and one or more climate indices may be considered together, the combination of multiple encoders may represent a valuable strategy to collect as much information as possible for the neural networks.

The final SSHa field is created by linearly combining the predicted PC values with the GCOOS decomposed EOF patterns, which are thought to be the proper dominant modes. The field rebuilt is in GCOOS space (Figure 10), and therefore lower resolution than the original HYCOM data. Although the focus of this study is on SSHa prediction with mesoscale features, which are captured by GCOOS, users may want the reconstructed field to be at the resolution of the analysis. For example, an additional downscaling process could be desirable for fields that contain the imprinting of submesoscale circulations such as temperature or salinity. For this purpose, we propose employing a SRCNN (Super Resolution Convolutional Neural Network) to convert the low-resolution field to its high-resolution counterpart. Super-resolution is a class of techniques that enhances the resolution of imaging system and has achieved great success in classical computer vision work (Dong et al., 2016). Recently, this technique has been applied to geoscience data set to recover the wind field at 1–4 km scale from a large-scale climate

model (Höhlein et al., 2020; Stengel et al., 2020) and for downscaling SSH fields (Barthélémy et al., 2022). We have adopted the SR technique and demonstrated that it yields nearly identical performance on sea surface level interpolation compared to the conventional cubic spline, corroborating the results by Barthélémy et al. (2022). We refer the reader to the Supporting Information S1 for detailed information on the implementation and performance of the SRCNN. It is worth mentioning that the SRCNN is essentially an image-based black-box model that lacks physical constraints. Therefore, we suggest using it where it outperforms traditional interpolation techniques after ensuring that the added noise does not deteriorate the outcome in terms of, for example, overall energy or spectral power distribution.

Data Availability Statement

All data sets analyzed in this work are publicly available. The GCOOS seas surface height data can be downloaded at https://geo.gcoos.org/ssh/ (last access: 09/17/2022). The HYCOM sea surface height data can be accessed through the https://www.hycom.org/dataserver (https://www.hycom.org/data/goml0pt04/expt-31pt0, and https://www.hycom.org/data/goml0pt04/expt-32pt5) (last access: 09/24/2022). The codes developed for the work (seq2seq and SRCNN) are available on zenodo (https://zenodo.org/doi/10.5281/zenodo.10070437).

Acknowledgments
AB and JB conceptualized this work during the KITP Program "Machine Learning and the Physics of Climate" supported by the National Science Foundation under Grant NSF PHY-1748958. This work is also funded by the National Oceanic and Atmospheric Administration (NOAA): NA18NOS478166. We thank three anonymous reviewers for their insightful comments and suggestions that greatly improved the manuscript.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459. https://doi.org/10.1002/WICS.101

Alvera-Azcárate, A., Barth, A., & Weisberg, R. H. (2009). The surface circulation of the Caribbean Sea and the Gulf of Mexico as inferred from satellite altimetry. *Journal of Physical Oceanography*, *39*(3), 640–657. https://doi.org/10.1175/2008JPO3765.1

Ballarotta, M., Ubelmann, C., Pujol, M. I., Taburet, G., Fournier, F., Legeais, J. F., et al. (2019). On the resolutions of ocean altimetry maps. *Ocean Science*, *15*(4), 1091–1109. https://doi.org/10.5194/OS-15-1091-2019

Barthélémy, S., Brajard, J., Bertino, L., & Counillon, F. (2022). Super-resolution data assimilation. *Ocean Dynamics*, *72*(8), 661–678. https://doi.org/10.1007/s10236-022-01523-x

Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, *45*(12), 6289–6298. https://doi.org/10.1029/2018gl078510

Chang, Y. L., & Oey, L. Y. (2013). Loop current growth and eddy shedding using models and observations: Numerical process experiments and satellite altimetry data. *Journal of Physical Oceanography*, *43*(3), 669–689. https://doi.org/10.1175/jpo-d-12-0139.1

Chapman, W. E., Subramanian, A. C., Delle Monache, L., Xie, S. P., & Ralph, F. M. (2019). Improving atmospheric river forecasts with machine learning. *Geophysical Research Letters*, *46*(17–18), 10627–10635. https://doi.org/10.1029/2019GL083662

Chattopadhyay, A., Nabizadeh, E., & Hassanzadeh, P. (2020). Analog forecasting of extreme-causing weather patterns using deep learning. *Journal of Advances in Modeling Earth Systems*, *12*(2), e2019MS001958. https://doi.org/10.1029/2019MS001958

Cho, K., van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST 2014—8th workshop on syntax, semantics and structure in statistical translation* (pp. 103–111). https://doi.org/10.48550/arxiv.1409.1259

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP 2014—2014 conference on empirical methods in natural language processing, proceedings of the conference* (pp. 1724–1734). https://doi.org/10.48550/arxiv.1406.1078

Cummings, J. A., & Smedstad, O. M. (2013). Variational data assimilation for the global ocean [Dataset]. Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II), 303–343. https://doi.org/10.1007/978-3-642-35088-7_13/COVER

Davò, F., Alessandrini, S., Sperati, S., Delle Monache, L., Airoldi, D., & Vespucci, M. T. (2016). Post-processing techniques and principal component analysis for regional wind power and solar irradiance forecasting. *Solar Energy*, *134*, 327–338. https://doi.org/10.1016/J.SOLENER.2016.04.049

Dong, C., Loy, C. C., He, K., & Tang, X. (2016). Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(2), 295–307. https://doi.org/10.1109/TPAMI.2015.2439281

Dueben, P. D., & Bauer, P. (2018). Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, *11*(10), 3999–4009. https://doi.org/10.5194/GMD-11-3999-2018

Falasca, F., & Bracco, A. (2022). Exploring the tropical Pacific manifold in models and observations. *Physical Review X*, *12*(2), 021054. https://doi.org/10.1103/PhysRevX.12.021054

Faranda, D., Messori, G., & Yiou, P. (2017). Dynamical proxies of North Atlantic predictability and extremes. *Scientific Reports*, *7*(1), 1–10. https://doi.org/10.1038/srep41278

Hall, C. A., & Leben, R. R. (2016). Observational evidence of seasonality in the timing of loop current eddy separation. *Dynamics of Atmospheres and Oceans*, *76*, 240–267. https://doi.org/10.1016/J.DYNATMOCE.2016.06.002

Hamilton, P. (1990). Deep currents in the Gulf of Mexico. *Journal of Physical Oceanography*, *20*(7), 1087–1104. https://doi.org/10.1175/1520-0485(1990)020<1087:DCITGO>2.0.CO;2

Herman, G. R., & Schumacher, R. S. (2018). Money doesn't grow on trees, but forecasts do: Forecasting extreme precipitation with random forests. *Monthly Weather Review*, *146*(5), 1571–1600. https://doi.org/10.1175/MWR-D-17-0250.1

Höhlein, K., Kern, M., Hewson, T., & Westermann, R. (2020). A comparative study of convolutional neural network models for wind field downscaling. *Meteorological Applications*, *27*(6), e1961. https://doi.org/10.1002/MET.1961

Jaimes, B., Shay, L. K., & Brewster, J. K. (2016). Observed air-sea interactions in tropical cyclone Isaac over Loop Current mesoscale eddy features. *Dynamics of Atmospheres and Oceans*, *76*, 306–324. https://doi.org/10.1016/j.dynatmoce.2016.03.001

Kingma, D. P., & Ba, J. L. (2014). Adam: A method for stochastic optimization. In *3rd international conference on learning representations, ICLR 2015—Conference track proceedings*. https://doi.org/10.48550/arxiv.1412.6980

Kurth, T., Treichler, S., Romero, J., Mudigonda, M., Luehr, N., Phillips, E., et al. (2019). Exascale deep learning for climate analytics. In *Proceedings—International conference for high performance computing, networking, storage, and analysis, SC 2018* (pp. 649–660). https://doi.org/10.1109/SC.2018.00054

Lagerquist, R., McGovern, A. M. Y., & Gagne, D. J. (2019). Deep learning for spatially explicit prediction of synoptic-scale fronts. *Weather and Forecasting*, *34*(4), 1137–1160. https://doi.org/10.1175/WAF-D-18-0183.1

Lamb, A. M., Alias Parth Goyal, A. G., Zhang, Y., Zhang, S., Courville, A. C., & Bengio, Y. (2016). Professor forcing: A new algorithm for training recurrent networks. *Advances in Neural Information Processing Systems*, *29*.

Larraondo, P. R., Renzullo, L. J., Inza, I., & Lozano, J. A. (2019). A data-driven approach to precipitation parameterizations using convolutional encoder-decoder neural networks. https://doi.org/10.48550/arxiv.1903.10274

Leben, R. R., Born, G. H., & Engebreth, B. R. (2010). Operational altimeter data processing for mesoscale monitoring [Dataset]. Marine Geodesy, *25*(1–2), 3–18. https://doi.org/10.1080/014904102753516697

Liu, G., Bracco, A., & Passow, U. (2018). The influence of mesoscale and submesoscale circulation on sinking particles in the northern Gulf of Mexico. *Elementa: Science of the Anthropocene*, *6*(1), 36. https://doi.org/10.1525/elementa.292

Liu, G., Bracco, A., & Sun, D. (2022). Offshore freshwater pathways in the northern Gulf of Mexico: Impacts of modeling choices. *Frontiers in Marine Science*, *9*, 841900. https://doi.org/10.3389/fmars.2022.841900

Liu, G., Falasca, F., & Bracco, A. (2021). Dynamical characterization of the loop current attractor. *Geophysical Research Letters*, *48*(24), e2021GL096731. https://doi.org/10.1029/2021GL096731

Liu, Y., MacFadyen, A., Ji, Z. G., & Weisberg, R. H. (2013). Monitoring and modeling the deepwater horizon oil spill: A record breaking enterprise. *Monitoring and Modeling the Deepwater Horizon Oil Spill: A Record Breaking Enterprise*, 1–271. https://doi.org/10.1029/GM195

Liu, Y., Weisberg, R. H., Vignudelli, S., & Mitchum, G. T. (2014). Evaluation of altimetry-derived surface current products using Lagrangian drifter trajectories in the eastern Gulf of Mexico. *Journal of Geophysical Research: Oceans*, *119*(5), 2827–2842. https://doi.org/10.1002/2013JC009710

Liu, Y., Weisberg, R. H., Vignudelli, S., & Mitchum, G. T. (2016). Patterns of the loop current system and regions of sea surface height variability in the eastern Gulf of Mexico revealed by the self-organizing maps. *Journal of Geophysical Research: Oceans*, *121*(4), 2347–2366. https://doi.org/10.1002/2015JC011493

Lorenz, E. N. (1980). Attractor sets and quasi-geostrophic equilibrium. *Journal of the Atmospheric Sciences*, *37*(8), 1685–1699. https://doi.org/10.1175/1520-0469(1980)037<1685:ASAQGE>2.0.CO;2

Lucarini, V., Faranda, D., de Freitas, J. M. M., Holland, M., Kuna, T., Nicol, M., et al. (2016). *Extremes and recurrence in dynamical systems*. John Wiley & Sons.

McGovern, A., Elmore, K. L., Gagne, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., et al. (2017). Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bulletin of the American Meteorological Society*, *98*(10), 2073–2090. https://doi.org/10.1175/BAMS-D-16-0123.1

NASEM, National Academies of Sciences, Engineering and Medicine. (2018). *Understanding and predicting the Gulf of Mexico loop current: Critical gaps and recommendations*. National Academies Press.

Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, *452*, 48–62. https://doi.org/10.1016/J.NEUCOM.2021.03.091

Rasp, S., & Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, *146*(11), 3885–3900. https://doi.org/10.1175/MWR-D-18-0187.1

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(39), 9684–9689. https://doi.org/10.1073/pnas.1810286115

Rodrigues, E. R., Oliveira, I., Cunha, R., & Netto, M. (2018). DeepDownscale: A deep learning strategy for high-resolution weather forecast. In *Proceedings—IEEE 14th international conference on EScience, e-Science 2018* (pp. 415–422). https://doi.org/10.1109/ESCIENCE.2018.00130

Scher, S., & Messori, G. (2019). Weather and climate forecasting with neural networks: Using general circulation models (GCMs) with different complexity as a study ground. *Geoscientific Model Development*, *12*(7), 2797–2809. https://doi.org/10.5194/GMD-12-2797-2019

Stengel, K., Glaws, A., Hettinger, D., & King, R. N. (2020). Adversarial super-resolution of climatological wind and solar data. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(29), 16805–16815. https://doi.org/10.1073/PNAS.1918964117/ASSET/A541AE37-AF00-46E8-96E6-E80275004FB2/ASSETS/IMAGES/LARGE/PNAS.1918964117FIG09

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, *27*.

Wang, J. L., Zhuang, H., Chérubin, L. M., Ibrahim, A. K., & Muhamed Ali, A. (2019). Medium-term forecasting of loop current eddy Cameron and eddy Darwin formation in the Gulf of Mexico with a divide-and-conquer machine learning approach. *Journal of Geophysical Research: Oceans*, *124*(8), 5586–5606. https://doi.org/10.1029/2019JC015172

Weisberg, R. H., & Liu, Y. (2017). On the loop current penetration into the Gulf of Mexico. *Journal of Geophysical Research: Oceans*, *122*(12), 9679–9694. https://doi.org/10.1002/2017JC013330

Weyn, J. A., Durran, D. R., & Caruana, R. (2019). Can machines learn to predict weather? Using deep learning to predict gridded 500-hPa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2680–2693. https://doi.org/10.1029/2019MS001705

Zanna, L., & Bolton, T. (2020). Data-driven equation discovery of ocean mesoscale closures. *Geophysical Research Letters*, *47*(17), e2020GL088376. https://doi.org/10.1029/2020gl088376